Ein verbandstheoretisches Modell zur Prognose von Kreditausfallwahrscheinlichkeiten

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Petra Fakler
aus Memmingen

Hundt Druck GmbH, Köln
2007

Berichterstatter: Prof. Dr. R. Schrader
Prof. Dr. E. Speckenmeyer

Tag der mündlichen Prüfung: 14.02.2007

Kurzzusammenfassung

Die Quantifzierung der Kreditrisiken und daraus resultierend die Mindesteigenkapitalanforderungen stellen im Bankenbereich eine zentrale Aufgabe dar. Aus Sicht von Basel II lassen sich durch adäquate Risikomessverfahren verringerte Eigenkapitalanforderungen erzielen. Wir stellen in der vorliegenden Arbeit ein Modell zur Ermittlung von Kreditausfallwahrscheinlichkeiten und zur Klassifikation von Darlehen vor. Die Modellierung basiert auf einem verbandstheoretischen Ansatz, der die Struktur in den verschiedenen Ausgangsklassen analysiert und es ermöglicht, aus diesen Strukturen ein interpretierbares Regelwerk zu erstellen. Diese Vorgehensweise liefert zudem eine natürliche Zerlegung des Darlehensbestandes in verschiedene Bonitätsklassen, wie sie bei Verwendung von internen Ratingverfahren aufsichtlich gefordert wird. Anhand von empirischen Wahrscheinlichkeiten, die aus realen Kollektivdaten ermittelt wurden, können die entstandenen Bonitätsklassen bewertet werden. Zur Validierung des Gesamtmodells wird die Problemstellung mit bekannten Klassifikationsverfahren bearbeitet und die Ergebnisse mit denen des verbandstheoretischen Implikationenmodells verglichen. Zudem wird die Generalisierungsfähigkeit aller untersuchten Modelle überprüft, indem die Ergebnisse auf ein zusätzliches reales Darlehenskollektiv übertragen werden. In einem letzten Schritt wird das verbandstheoretische Implikationenmodell auf weitere bauspartechnische Fragestellungen angewendet.

Abstract

Quantifiying the risk of credits and satisfying the resulting minimum capital standards is one of the major challenges in credit banking. In the view of Basel II, it is possible to reduce the minimum capital requirements by the use of apropriate assessments of risk. In the work at hand, we present a model to compute the probability of default and to classify loans. The model is based on a lattice theoretic approach, which analyzes the structure of properties in different credit classes and permits the creation of an interpretable body of rules. Furthermore, this procedure generates an elementary decomposition of loans into different borrower grades, which is necessary when the internal ratings–based approach (IRB–Approach) is used. Using empirical probabilities, acquired from real–world data, it is possible to assess the resulting borrower grades. To validate the whole model, the problem is also solved using different established classification models, and results are compared with the lattice theoretic approach. Furthermore, the generalization ability of all models is examined by transferring the models to another loan collective. Finally, the lattice theory based model is tested on some other building association specific problems.

Vielen Dank!

Ich möchte Herrn Prof. Dr. Faigle und Herrn Prof. Dr. Schrader danken, dass ich in Ihrer Arbeitsgruppe am Zentrum für Angewandte Informatik Köln (ZAIK) arbeiten durfte. Mein besonderer Dank gilt dabei Herrn Prof. Dr. Schrader, der diese Arbeit betreute. Die anregenden Gespräche und Diskussionen trugen maßgeblich zum Gelingen der Arbeit bei. Weiterhin möchte ich mich bei Herrn Prof. Dr. Speckenmeyer für die Übernahme des Korreferats bedanken.

Mein Dank richtet sich auch an die Gruppe der Landesbausparkassen, die das Projekt finanzierten, in dessen Rahmen diese Arbeit entstanden ist. Zudem unterstützten sie diese Arbeit mit zahlreichen Anregungen und persönlichem Engagement.

Weiterhin möchte ich meinen Kollegen Dr. Thomas Chevalier, Martin Lätsch und Stefan Neuhaus aus der Bausparkassengruppe danken, die sich die Mühe machten, die Arbeit zu lesen und mich auf Fehler hinzuweisen. Die gute Atmosphäre in der Gruppe war eine tägliche Motivation zur Erstellung dieser Arbeit. Vielen Dank!

Ich danke meinen ehemaligen und aktuellen Kolleginnen und Kollegen Bernhard Fuchs, Dr. Christian Hagemeier, Dr. Markus Kloock, Dr. Dirk Räbiger, Dr. Britta Peis und einigen anderen für die schöne Zeit am ZAIK.

Mein Dank gebührt auch meiner Familie. Meinen Eltern Willebold und Diana Fakler, die mich auf meinem bisherigen Lebensweg immer unterstützt haben. Meinem Bruder Thomas und Brigitte, Peter und Andreas Fakler für die gebotene Ablenkung während meiner Zeit am ZAIK.

Ganz besonderer Dank gebührt meinem Freund Michael Grimmer, der mich mit seiner positiven Energie und Lebenseinstellung immer wieder aufgemuntert und motiviert hat.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Kapitel 1

Einleitung

Ab dem Jahr 2007 gelten für alle europäischen Kreditinstitute die neuen Eigenkapitalrichtlinien Basel II. Damit sind vor allem Neuerungen bei der Risikoeinstufung von Krediten und der damit verbundenen Eigenkapitalhinterlegung betroffen.
Kreditinstitute benötigen Eigenkapital um die Verluste aus eingetretenen Markt- und Kreditrisiken auszugleichen. Eine wichtige Forderung ist daher, dass die Kreditinstitute ihre Risiken angemessen durch Eigenkapital hinterlegen. Die Einführung der neuen Eigenkapitalrichtlinie wird vor allem mit einer Stabilisierung und Sicherung des Bankensystems begründet.
Bereits Ende der 80er Jahre wurde die Eigenkapitalrichtlinie Basel I verabschiedet. Danach soll das Eigenkapital mindestens 8% der gewichteten Risikoaktiva[1] betragen [MA04]. Die Risikomessung erfolgte allerdings bislang pauschal in vier Risikoklassen mit vorgegebenen Risikogewichten $(0\%, 20\%, 50\%$ und $100\%)$. Alle Kredite an Privatpersonen und Privatunternehmen wurden ausnahmslos mit 100% bewertet, also mit einer Eigenkapitalquote von 8% belastet.
Dieser pauschale Ansatz steht jedoch im Widerspruch zu der sich verändernden Situation in den Kreditinstituten. Die bankinternen Methoden zur Messung des betriebswirtschaftlichen Eigenkapitals und die Messung der Kreditrisiken haben sich erheblich weiterentwickelt [CDEL05]. Ein pauschales Risikogewicht für alle Privatpersonen und Privatunternehmen scheint daher nicht mehr zeitgemäß. Zudem wurden Methoden zur Risikominderung von Basel I nur wenig anerkannt.

Daher begann der Baseler Ausschuss für Bankenaufsicht im Jahre 1999 damit, eine neue risikogerechte Regelung der Eigenkapitalhinterlegung zu formulieren. Die neue Eigenkapitalvereinbarung (Basel II) beruht auf drei Säulen: der Regelung der Mindesteigenkapitalanforderungen, der Überprüfung dieser Anforderungen durch Aufsichtsbehörden sowie der Transparenz und Marktdisziplin, welche die Offenlegungspflichten der Banken umfassen [Cre04]. Die grundlegende Eigenkapitaldefinition und der Eigenkapitalkoeffizient bleiben unter Basel II unverändert. Das innovativste Element

[1] Basel I umfasste dabei Kredit- und Marktrisiken.

der neuen Eigenkapitalvereinbarung ist die genauere Messung der Bonität eines Kreditnehmers. Damit wird das Risiko bei der Kreditvergabe transparenter und zukünftige Kreditverluste verringert bzw. können bei Kreditvergabe entsprechend bepreist werden [Meu03]. Mit Hilfe der risikogerechten Beurteilung von Kreditnehmern lassen sich für Kreditinstitute geringere Eigenkapitalhinterlegungen erzielen.
Zudem gestattet Basel II die Verwendung von externen Ratings zur Bonitätsbeurteilung der Kreditnehmer. Dadurch können Kreditinstitute auf externe Bonitätsbeurteilungsinstitutionen (z. B. Standard & Poor's oder Moody's) zurückgreifen oder die Ausfallwahrscheinlichkeiten der Kreditnehmer, im Rahmen sogenannter IRB[2]–Ansätze intern ermitteln.
Nach den Beschlüssen des Baseler Ausschusses für Bankenaufsicht stellt die Bonitätseinstufung der Kreditnehmer durch eine externe Ratingagentur oder interne Ratings in Zukunft das wesentliche Kriterium dar, nach der die Eigenkapitalhinterlegung ermittelt wird. Damit sind allerdings auch die Kreditinstitute gefordert, ihre Kreditnehmer zu bewerten. Die Ermittlung der entsprechenden Risikogewichte erfolgt unter anderem mit Hilfe der Ausfallwahrscheinlichkeiten und dem zu erwarteten Verlust bei Ausfall.
Damit sind auch Bausparkassen verpflichtet Bonitätsbeurteilungen für ihre Kunden durchzuführen. Allerdings können Bausparkassen nicht auf externe Ratings zurückgreifen, da die Ratingagenturen meist aus dem amerikanischen Raum stammen und nicht mit dem System des Bausparens vertraut sind und somit keine Bewertungen vorliegen. Dies gilt ebenso für deutsche mittelständische Unternehmen [BCK04]. Daher sind die Bausparkassen gezwungen auf IRB–Ansätze und damit auf interne Bewertungen zurückzugreifen. Für die Bausparkassen, die größtenteils Baukredite an Privatkunden vergeben, sind die Voraussetzungen aber günstig, dass mit internen Bonitätsbeurteilungen die Kreditrisiken des Bestandes zuverlässig ermittelt werden können [Zeh01].

Da die Bausparkassen auf interne Ratings angewiesen sind, wird in dieser Arbeit ein logisches Modell zur Klassifikation von Bauspardarlehen und zur Ermittlung von Kreditausfallwahrscheinlichkeiten vorgestellt. Die Kreditausfallwahrscheinlichkeiten werden zur Ermittlung der Risikogewichte und zur Bildung von Bonitätsklassen benötigt. Das Modell soll den Anwender bei der Beurteilung der Kreditnehmer unterstützen und ihm dabei ein überschaubares und nachvollziehbares Regelwerk zur Bewertung liefern.
Um die Qualität der ermittelten Ausfallwahrscheinlichkeiten und die Klassifikationsgüte zu überprüfen, werden die Ergebnisse mit anderen logischen sowie probabilistischen Modellen verglichen. Zur weiteren Validierung wird das Modell auf eine zusätzliche Bausparkasse übertragen. Damit kann die Generalisierungsfähigkeit der Modellierung bewertet werden. In einem weiteren Schritt wird die Anwendung des logischen Modells auf andere bauspartechnische Fragestellungen ausgeweitet. Hierbei ist vor al-

[2] IRB = auf internen Ratings basierender Ansatz (Internal Ratings Based Approach).

lem die Klassifikation von Darlehensverzichtern bzw. Darlehensnehmern zu erwähnen. Daraus ergibt sich für die vorliegende Arbeit folgende Struktur:
In Kapitel 2 werden die Modellgrundlagen vorgestellt. Dabei wird zu Beginn auf das Klassifikationsproblem im Allgemeinen eingegangen. Weiterhin werden notwendige mathematische und bauspartechnische Grundlagen zur Modellierung dargestellt.
Da die Klassifikationsergebnisse mit den Ergebnissen anderer Modelle verglichen werden sollen, werden in Kapitel 3 die theoretischen Grundlagen drei weiterer Klassifikationsmodelle beschrieben. Dabei handelt es sich um zwei logische und ein probabilistisches Modell.
Ziel des Kapitels 4 ist die Entwicklung eines Modells zur Klassifizierung und Bewertung von Bauspardarlehen. Da die Modellierung auf einem verbandstheoretischen Ansatz basiert, wird das Modell auch als verbandstheoretisches Implikationenmodell bezeichnet. Ausgehend von der Datendiskretisierung, die für logische Modelle notwendig ist, wird die Ermittlung eines signifikanten Regelwerks zur Klassifikation beschrieben. Anschließend wird auf die Bewertung des Regelwerks und die Analyse der Ergebnisse eingegangen.
In Kapitel 5 wird das verbandstheoretische Implikationenmodell auf reale Kollektivdaten angewendet um Kreditausfälle zu klassifizieren und die Ausfallwahrscheinlichkeiten der Kreditnehmer zu ermitteln. Die zugrunde liegende Bausparkasse werden wir im Folgenden als Ausgangsbausparkasse bezeichnen. Zusätzlich wird das Modell ohne weitere Anpassungen auf die Originaldaten einer weiteren Bausparkasse angewendet und die gewonnenen Ergebnisse dargestellt. Diese Bausparkasse werden wir in den folgenden Ausführungen als Validierungsbausparkasse bezeichnen. Abschließend wird auf eine mögliche Verwendung der Ergebnisse im Rahmen von IRB–Ansätzen eingegangen.
Da die Ergebnisse des verbandstheoretischen Implikationenmodells mit anderen Modellen verglichen werden sollen, werden in Kapitel 6 die Klassifikationsergebnisse der in Kapitel 3 vorgestellten Modelle dargestellt. Auch diese Modelle werden zur Überprüfung ihrer Generalisierungsfähigkeit ohne weitere Anpassung auf die Validierungsbausparkasse übertragen.
Ein quantitativer und qualitativer Vergleich aller Ergebnisse erfolgt in Kapitel 7. Um die Vergleichbarkeit der verschiedenen Modelle zu gewährleisten, wurde eine einheitliche Diskretisierung der logischen Modelle durchgeführt. Im ersten Teil des Kapitels wird ein quantitativer Vergleich der Klassifikationsergebnisse durchgeführt. Im zweiten Teil werden die Unterschiede der Modelle inhaltlich erläutert und interpretiert. Dazu wurden weitere Datenanalysen durchgeführt.
Abschließend wird das verbandstheoretische Implikationenmodell in Kapitel 8 auf weitere bauspartechnische Fragestellungen angewandt. Dabei wurde die Struktur der Darlehensverzichter bzw. Darlehensnehmer untersucht und eine Klassifikation vorgenommen. Eine weitere Klassifikation wurde bei den Kündigern in Sperrfrist durchgeführt. Kapitel 9 bewertet abschließend die Ergebnisse nochmals und geht auf mögliche Verbesserungen und Erweiterungen des Modells ein.

Teil I

Theorie

Kapitel 2

Modellgrundlagen

2.1 Klassifikation und Klassenbewertung

Das Problem der Zuordnung eines Objektes zu einer von mehreren gegebenen Klassen wird als Klassifikation bezeichnet. Beispiele für Klassifikationsprobleme sind die Zuordnung von kranken und gesunden Patienten, die Einordnung von Darlehensnehmer in verschiedene Risikoklassen, oder die Zuordnung von Tieren bzw. Pflanzen in bestimmte Gattungen. Dabei sind die Anwendungsmöglichkeiten außerordentlich vielseitig.
Damit eine solche Zuordnung zu verschiedenen Klassen möglich ist, werden Informationen über die Objekte benötigt. Die Objekte werden in den folgenden Ausführungen auch als Trainingsbeispiele bezeichnet. Die Informationen über die Objekte liegen in Form von Merkmalen bzw. Merkmalsausprägungen vor. Klassifikationsmodelle versuchen mit Hilfe der in den Trainingsbeispielen enthaltenen Informationen die Zuordnung zur jeweiligen Klasse bzw. eine Trennung der Klassen vorzunehmen. Dabei soll die Zuordnung binär sein, d. h. ein Objekt kann genau einer Klasse zugeordnet werden. Allerdings stellt sich die Frage, welche Merkmale für die Trennung bzw. Klassenzuordnung verwendet werden sollen. Meist enthalten die Objekte eine Vielzahl von Merkmalen. Welche dieser Merkmale können sinnvoll zur Klassifikation verwendet werden? Es ist zudem möglich, dass nur bestimmte Merkmalskombinationen für die Klassenzugehörigkeit verantwortlich sind, d. h. einzelne Merkmale können redundant sein. Wurde eine Merkmalsauswahl getroffen, so sollte diese Auswahl auch in der Lage sein, neue Objekte mit denselben Merkmalen sinnvoll zu klassifizieren.

Werden bei der Klassifikation unendlich viele Klassen zugelassen, so handelt es sich um ein Regressionsproblem. Dabei wird einem Objekt ein reeller Wert zugeordnet. Die Regression stellt damit eine Verallgemeinerung des Klassifikationsproblems dar. Beispiele aus dem ökonomischen Bereich können z. B. Waren sein, die durch ihren Preis, ihre Verkaufsförderung und Rabattaktionen beschrieben werden. Diesen Waren werden die entsprechenden Absatzmengen zugeordnet. Im Folgenden werden wir uns allerdings auf Klassifikationsprobleme mit endlicher Klassenanzahl

beschränken. Weitere Informationen zur Regressionsanalyse finden sich unter anderem in [DHS01, BLK06, BEP+86, OU02, Kad06].

Im Verlauf dieser Arbeit wird aber nicht nur die Zuordnung der Trainingsbeispiele zu den Klassen, sondern auch die Bewertung der Klassen von Bedeutung sein. Diese Klassenbewertung erlaubt eine zusätzliche Verwendung der Klassifikationsergebnisse, die in Unterabschnitt 2.7 vorgestellt wird.

2.2 Klassifikationsproblem

2.2.1 Überwachtes/Nicht überwachtes Lernen

Beim Lernen wird generell zwischen dem überwachten und dem nicht überwachten Lernen unterschieden [RN04]. Das überwachte Lernen zeichnet sich durch die vorab bekannte Klassenzugehörigkeit der Objekte aus. Dem zu lernenden System wird mitgeteilt, aus welcher Klasse ein Objekt stammt. Beim überwachten Lernen wird also stets mit Hilfe von bekannten Beispielen eine Klassifizierungsfunktion gelernt. Ein- und Ausgabewerte beim überwachten Lernen sind vorab bekannt. Klassische Verfahren des überwachten Lernens sind neuronale Netze, Entscheidungsbäume und Regressionsverfahren.

Beim nicht überwachten Lernen liegen Trainingsbeispiele mit unbekannter Klassenzugehörigkeit vor. Dies wirft die Frage auf, wie Objekte mit unterschiedlichen Eigenschaften in sinnvolle Kategorien eingeteilt, und wie solche Kategorien entdeckt werden können. Ein Beispiel für ein Kategorisierungsproblem ist die Einteilung von Bausparern zu ähnlichen Sparertypen [Van96]. Zur Kategorisierung können unter anderem Clusterverfahren verwendet werden. Eine Darstellung weiterer Verfahren findet sich unter anderem in [BEP+86].

Aufgrund der bekannten Klassenzugehörigkeit in den vorliegenden Trainingsbeispielen, werden wir uns in den folgenden Ausführungen auf das überwachte Lernen beschränken. Im Verlauf dieser Arbeit wollen wir unsere Problemstellung mit unterschiedlichen Klassifikationsverfahren modellieren.

2.2.2 Mathematische Formulierung

Im folgenden Abschnitt wollen wir auf die mathematische Formulierung des überwachten Lernproblems, wie sie z. B. bei der Modellierung mit neuronalen Netzen benötigt wird, näher eingehen. Aus den bereits klassifizierten Objekten $X = (x_1, \ldots, x_n)$ soll eine Klassifikationsfunktion h ermittelt werden, die allen Objekten die jeweilige Klasse y zuordnet. Dabei sei $\mathbb{F}$ der Merkmalsraum, in dem alle zur Klassifikation

relevanten Objekte beschrieben sind. Da viele Merkmale reellwertige Einträge enthalten, wird $\mathbb{F} = \mathbb{R}^n$ gesetzt. Zusätzlich zum Merkmalsraum $\mathbb{F}$ ist ein Ergebnisraum $\mathbb{L}$ mit zugehörigen Klassen gegeben. Beim Klassifikationsproblem enthält $\mathbb{L}$ nur diskrete Werte, nämlich die Zahl der Klassennummern, und damit gilt $\mathbb{L} \subset \mathbb{N}$. Es ist eine Funktion $h(x) = y$ gesucht, die jedem Objekt bzw. dem Merkmalsvektor x der dieses Objekt beschreibt, seine Klasse y zuordnet. Die Bezeichnung x kennzeichnet also das Objekt sowie den objektbeschreibenden Merkmalsvektor. Ziel ist es, auch für zukünftige Testobjekte gute Vorhersagen zur erreichen. Dabei entspricht $h(x)$ der idealen Zielfunktion der Trainingsbeispiele. Allerdings ist es im Allgemeinen sehr schwierig eine solche Zielfunktion perfekt zu lernen. Häufig sind die Trainingsbeispiele z. B. durch Messfehler verrauscht. Daher wird versucht, diese ideale Zielfunktion möglichst gut zu approximieren. Der Lernprozess wird häufig als Funktionenapproximation bezeichnet [Mit97]. Die gesuchte Funktionenapproximation bezeichnen wir mit $\hat{h}(x)$, um diese von der idealen Zielfunktion $h(x)$ zu unterscheiden. Diese Approximation beinhaltet natürlicherweise auch Fehlklassifikationen, d. h. ein Objekt mit bekannter Klassenzugehörigkeit wird fälschlicherweise einer anderen Klasse zugeordnet. Ein weiteres Ziel ist daher die Minimierung der Fehlklassifikationen.
Bei parametrischen Verfahren unterstellt man dem zu modellierenden Zusammenhang eine bestimmte funktionale Form. Anschließend bestimmt man die Parameter der Funktion mit Hilfe der vorliegenden Trainingsbeispiele. Ein einfaches Beispiel ist die lineare Regression. Bei Vorlage der Daten in Form (y_i, x_i), $i = 1, \ldots, n$ wird dem Modell ein linearer Zusammenhang der Form $Y_1 = \beta_0 + \beta_1 x_1$ mit $(\beta_0, \beta_1) \in \mathbb{R}^n$ unterstellt. Die Parameter β_0 und β_1 können anschließend z. B. mit der Methode der kleinsten Fehlerquadrate ermittelt werden.

2.2.3 Aussagenlogische Formulierung

Da im Rahmen dieser Arbeit logische Klassifikationsmodelle verwendet werden, wird an dieser Stelle auf die aussagenlogische Formulierung der Problemstellung eingegangen. Dabei werden wir kurz die Grundzüge der Aussagenlogik darstellen.

2.2.3.1 Aussagenlogische Ausdrücke

In der Aussagenlogik werden Ausdrücke betrachtet, die aus den Aussagenvariablen, den sogenannten Atomen, allein durch Hilfe von Junktoren zusammengesetzt sind. Die Aussagenvariablen werden durch ihre Wahrheitswerte „Wahr" und „Falsch" interpretiert. Wir wollen im Folgenden die konkrete Menge aussagenlogischer Ausdrücke (Formeln) definieren [EFT92]:

Definition 2.1. *Es sei $\mathcal{A}$ das Alphabet der Aussagenlogik bestehend aus*

- *einer endlichen Anzahl von Aussagenvariablen $x \in V$, mit $V = \{x_0, x_1, x_2, \ldots, x_n\}$, $n \in \mathbb{N}$ und*

- *Junktoren J=*$\{\top, \bot, \neg, \wedge, \vee, \rightarrow, \leftrightarrow\}$.

Für die aussagenlogischen Ausdrücke, als Zeichenreihen über $\mathcal{A}$ *gilt:*

- *Jedes Atom ist eine Formel, d. h. alle* $x \in V$ *sind Formeln,*
- $\bot$ *und* $\top$ *sind Formeln,*
- *sind A und B Formeln, dann sind auch* $(\neg A)$, $(A \wedge B)$, $(A \vee B)$, $(A \rightarrow B)$ *und* $(A \leftrightarrow B)$ *Formeln.*

Die aussagenlogischen Ausdrücke sind mit den Buchstaben $A, B, \ldots$ gekennzeichnet und setzen sich aus Aussagenvariablen und Junktoren zusammen. Dabei werden die Symbole der Menge J als „wahr", „falsch", „nicht", „und", „oder", „wenn–so" („impliziert") und „genau dann wenn" genannt. Die Menge aller aussagenlogischen Ausdrücke bezeichnen wir mit AA.
Ein Ausdruck ist dabei in disjunktiver Normalform (DNF), wenn er eine Disjunktion von Konjunktionen aus Aussagenvariablen oder negierten Aussagenvariablen ist, er ist in konjunktiver Normalform (KNF), wenn er eine Konjunktion von Disjunktionen aus Aussagenvariablen oder negierten Aussagenvariablen ist.

Beispiel 2.1. *Der Ausdruck*

$$(x_0 \vee x_3 \vee (\neg x_7 \wedge x_4))$$

ist in disjunktiver Normalform, der Ausdruck

$$((x_1 \vee \neg x_3) \wedge x_3 \wedge (\neg x_8 \vee x_2))$$

ist in konjunktiver Normalform.

2.2.3.2 Hornausdrücke und Problemformulierung

Im vorigen Abschnitt haben wir die Syntax der Aussagenlogik erläutert und die Menge der aussagenlogischen Ausdrücke AA definiert. In den folgenden Ausführungen wollen wir auf aussagenlogische Ausdrücke von besonderer Gestalt eingehen und unsere Problemstellung formulieren.

Definition 2.2. *[HW91] Ein aussagenlogischer Hornausdruck besitzt eine der folgenden Gestalten:*

1. $\neg x_1 \vee \neg x_2 \vee \ldots \vee \neg x_n \vee x$ *für* $n \in \mathbb{N}$
2. x
3. $\neg x_1 \vee \neg x_2 \vee \ldots \vee \neg x_n$ *für* $n \in \mathbb{N}$

Aussagenlogische Hornausdrücke besitzen also höchstens ein positives Literal. Weiterhin haben sie die Eigenschaft, dass man die Erfüllbarkeit solcher aussagenlogischen Hornausdrücke leicht testen kann. Mehr zu Hornausdrücken und dem Erfüllbarkeitsproblem findet sich in [EFT92].
Der Ausdruck in 1. ist aussagenlogisch äquivalent zu $x_1 \wedge x_2 \wedge \ldots \wedge x_n \rightarrow x$ $(AH1)$, wie man mit Hilfe von Wahrheitstabellen leicht zeigen kann. Diesen Ausdruck bezeichnen wir im Folgenden auch als Implikation.
Unsere Problemstellung lässt sich damit folgendermaßen formulieren. Gegeben sei eine (m,n)–Matrix mit m Bauspardarlehen und n Aussagenvariablen. Gesucht sind Merkmale bzw. Merkmalskombinationen, die für die Klassifikation von Bedeutung sind. Um die notwendigen Informationen aus den Trainingsbeispielen zu erlangen, werden wir in den folgenden Ausführungen Ausdrücke der Gestalt $AH1$ verwenden, da sie im Rahmen der formalen Begriffsanalyse in der Lage sind, die Struktur in den Trainingsbeispielen zu beschreiben. Eine ausführliche Beschreibung der Implikationen im Sinne der formalen Begriffsanalyse wollen wir in den folgenden Ausführungen darstellen.

2.3 Formale Begriffsanalyse und Implikationen

Gegenstand der formalen Begriffsanalyse sind binäre Relationen und mit ihnen in enger Verbindung stehend vollständige Verbände. Die formale Begriffsanalyse liefert zudem Anwendungen, um Zusammenhänge in Daten zu beschreiben und dadurch zu neuen Einsichten zu gelangen. Dabei ist unter anderem die Implikationentheorie von großer Bedeutung, da sich die Struktur eines Begriffsverbandes durch die ihn im geltenden Implikationen beschreiben lässt [Gan00].
Wir wollen daher in der vorliegenden Arbeit die Implikationentheorie nutzen, um die Struktur in unseren Trainingsbeispielen zu beschreiben. Mit Hilfe einer handlichen Menge von Implikationen wollen wir ein Regelwerk erzeugen, das in der Lage ist, unser Klassifikationsproblem zu lösen. Vorab wollen wir aber auf grundlegende Definitionen und Sätze der formalen Begriffsanalyse eingehen.

2.3.1 Hüllenoperatoren, Hüllensysteme

Definition 2.3. Hüllensystem, Hüllenoperator
Ein Hüllensystem auf einer Menge G ist eine Menge von Teilmengen, die G enthält und unter Durchschnittsbildung abgeschlossen ist. Formal: $\mathcal{U} \subseteq \mathcal{P}(G)$ ist ein Hüllensystem, falls $G \in \mathcal{U}$ ist und

$$\mathcal{X} \subseteq \mathcal{U} \Rightarrow \bigcap \mathcal{X} \in \mathcal{U}$$

gilt. Ein Hüllenoperator φ auf G ist eine Abbildung, die jeder Teilmenge $X \subseteq G$ eine Hülle $\varphi X \subseteq \mathcal{U}$ zuordnet, wobei gilt:

- $X \subseteq Y \Rightarrow \varphi X \subseteq \varphi Y$ *(Monotonie)*
- $X \subseteq \varphi X$ *(Extensität)*
- $\varphi\varphi X = \varphi X$ *(Idempotenz).*

Hüllensystem und Hüllenoperator sind eng miteinander verwandt, wie der folgende Satz zeigt:

Satz 2.1. *[GW96] Ist $\mathcal{U}$ ein Hüllensystem auf G, so definiert*

$$_{\varphi\mathcal{U}}X := \bigcap\{A \in \mathcal{U} | X \subseteq A\}$$

einen Hüllenoperator auf G. Umgekehrt ist die Menge

$$\mathcal{U}_\varphi := \{\varphi X | X \subseteq G\}$$

aller Hüllen eines Hüllenoperators φ stets ein Hüllensystem und es gilt:

$$\varphi\mathcal{U}_\varphi = \varphi \quad \textit{sowie} \quad \mathcal{U}_{\varphi\mathcal{U}} = \mathcal{U}$$

Beweis. - $\varphi\mathcal{U}$ ist ein Hüllenoperator: Aus $X \subseteq Y$ folgt

$$\{A \in \mathcal{U} \mid X \subseteq A\} \supseteq \{A \in \mathcal{U} \mid Y \subseteq A\},$$

also aufgrund der Monotonie von $\cap$

$$_{\varphi\mathcal{U}}X = \bigcap\{A \in \mathcal{U} \mid X \subseteq A\} \subseteq \bigcap\{A \in \mathcal{U} \mid Y \subseteq A\} = {}_{\varphi\mathcal{U}}Y.$$

Die Extensität ist trivial, Idempotenz: Nach der Definition von $\varphi\mathcal{U}$ enthält jedes Element von $\mathcal{U}$, welches X enthält, auch $_{\varphi\mathcal{U}}X$ und umgekehrt.

- $\mathcal{U}_\varphi$ ist ein Hüllensystem: Sei $\mathcal{X} \subseteq \mathcal{U}_\varphi$. Wegen der Extensität von φ ist $\bigcap\mathcal{X} \subseteq \varphi(\bigcap\mathcal{X})$. Mit Monotonie und Idempotenz folgt aus $X \in \mathcal{X}$ stets $\varphi(\bigcap\mathcal{X}) \subseteq \varphi X = X$, was $\varphi(\bigcap\mathcal{X}) \subseteq \bigcap\mathcal{X}$ nach sich zieht.

- $X \in \mathcal{U} \Leftrightarrow X = \bigcap\{A \in \mathcal{U} \mid X \subseteq A\} \Leftrightarrow {}_{\varphi\mathcal{U}}X = X \Leftrightarrow X \in \mathcal{U}_{\varphi\mathcal{U}}$.

- Für $A \in \mathcal{U}_\varphi$ ist $X \subseteq A$ gleichbedeutend zu $\varphi X \subseteq A$. Also ist $_{\varphi\mathcal{U}_\varphi}X = \bigcap\{A \in \mathcal{U}_\varphi \mid X \subseteq A\} = \bigcap\{A \in \mathcal{U}_\varphi \mid \varphi X \subseteq A\} = \varphi X$, da $\varphi X \in \mathcal{U}_\varphi$. □

Beispiele für Hüllensysteme sind z. B. die Potenzmenge, Untergruppen und auch der im folgenden Abschnitt eingeführte (Ableitungs–)Operator $''$ stellt einen Hüllenoperator mit zugehörigem Hüllensystem dar. Dieser ist für uns von besonderem Interesse, da er zur Ermittlung der Implikationen verwendet wird. Weitere Beispiele für Hüllenoperatoren mit zugehörigen Hüllensystemen finden sich in [GW96].

2.3.2 Kontext und Begriff

Die elementaren Grundbegriffe der formalen Begriffsanalyse sind formale Kontexte und formale Begriffe. Diese wollen wir im folgenden Abschnitt definieren.

Definition 2.4. Formaler Kontext
Ein formaler Kontext $\mathbb{K} := (G, M, I)$ besteht aus zwei Mengen G und M sowie einer Relation I zwischen G und M. Die Elemente von G nennen wir Gegenstände, die von M Merkmale. Um auszudrücken, dass ein Gegenstand g mit einem Merkmal m in Relation steht, schreiben wir gIm oder $(g,m) \in I$ und lesen dies als: der Gegenstand g besitzt Merkmal m.

Dabei bezeichnet I die Inzidenzrelation des Kontextes, $(g,m) \notin I \cong g \not I m$.
Kleinere Kontexte lassen sich bequem als „Kreuztabellen" darstellen. Ein Kreuz in Zeile g und Spalte m gibt an, dass der Gegenstand g das Merkmal m besitzt. Ein Kontext definiert, in welcher Beziehung eine Menge von Gegenständen zu einer Menge von Merkmalen stehen. Tabelle 2.1 zeigt beispielhaft einen Kontext über Planeten unseres Sonnensystems. Betrachtet man die Gegenstände g des Kontextes, so kann man

	klein	mittel	groß	nah	entfernt	Mond	kein Mond
Merkur	X			X			X
Venus	X			X			X
Erde	X			X		X	
Mars	X			X		X	
Jupiter			X		X	X	
Saturn			X		X	X	
Uranus		X			X	X	
Neptun		X			X	X	

Tabelle 2.1: Beispielkontext $\mathbb{K}$ über Planeten unseres Sonnensystems [Lin99].

nach deren gemeinsamen Merkmalen fragen. Analog kann man die gemeinsamen Gegenstände eines Merkmals m betrachten. Dazu definieren wir:

Definition 2.5. Ableitungsoperator $'$
Für eine Menge $A \subseteq G$ von Gegenständen definieren wir:

$$A' := \{m \in M \mid gIm \text{ für alle } g \in A\}$$

(die Menge der gemeinsamen Merkmale der Gegenstände in A). Entsprechend ist für eine Menge $B \subseteq M$ von Merkmalen

$$B' := \{g \in G \mid gIm \text{ für alle } m \in B\}$$

definiert (die Menge der Gegenstände, die alle Merkmale aus B besitzen).

Für unser Beispiel sind gemeinsame Merkmale von $g = \{Venus, Merkur\}$ die Menge $g' = \{klein, nah, kein\ Mond\}$, die Merkmale $m = \{gross, entfernt, Mond\}$ besitzen die Gegenstände $m' = \{Jupiter, Saturn\}$ alle gemeinsam.
Die zentrale Definition der formalen Begriffsanalyse ist die des formalen Begriffs:

Definition 2.6. Formaler Begriff
Ein formaler Begriff des Kontextes (G,M,I) ist ein Paar (A,B) mit $A \subseteq G,\ B \subseteq M,\ A' = B$ und $B' = A$. Wir nennen A den Umfang des Begriffes, und B den Inhalt des Begriffes (A,B). $\mathcal{B}(G,M,I)$ bezeichnet die Menge aller Begriffe des Kontextes (G,M,I).

Ein Begriff umfasst anschaulich ein maximales Rechteck im Kontext. In unserem Beispiel wäre $\{(Jupiter, Saturn, Uranus, Neptun), (entfernt, Mond)\}$ ein Begriff. Dieses „voll belegte" Rechteck ist in Tabelle 2.1 eingezeichnet.

Ein Begriff ist also ein Paar aus der Gegenstands- und Merkmalsmenge. Die Gegenstandsmenge eines Begriffes wird als Begriffsumfang, die Merkmalsmenge als Begriffsinhalt bezeichnet. Die Operation $'$ zur Bestimmung der gemeinsamen Gegenstände und Merkmale besitzt eine Reihe von Eigenschaften [GW96]:

Hilfssatz 2.2. *Ist (G,M,I) ein Kontext und sind $A, A_1, A_2 \subseteq G$ Mengen von Gegenständen und $B \subseteq M$ eine Menge von Merkmalen, so gilt*

1. $A_1 \subseteq A_2 \Rightarrow A_2' \subseteq A_1'$
2. $A \subseteq A''$
3. $A' = A'''$
4. $A \subseteq B' \iff B \subseteq A' \iff A \times B \subseteq I$.

Die analogen Eigenschaften (1.–3.) gelten natürlich auch für die Mengen von Merkmalen $B, B_1, B_2 \subseteq M$.

Beweis.
1. Ist $m \in A_2'$, so gilt gIm für alle $g \in A_2$, also erst recht gIm für alle $g \in A_1$, falls $A_1 \subseteq A_2$, und damit $m \in A_1'$.
2. Ist $g \in A$, so gilt gIm für alle $m \in A'$, woraus $g \in A''$ folgt.
3. $A \subseteq A''$ folgt aus der Eigenschaft $B \subseteq B''$, und aus $A \subseteq A''$ erhält man mit (1) $A''' \subseteq A'$.
4. folgt direkt aus der Definition.

□

Wendet man die Operation $'$ zweimalig an, so erhält man einen Hüllenoperator, wie er in Definition 2.3 beschrieben ist. Die Abbildung $\varphi : X \to X, \varphi(X) = X''$ ist monoton (1), extensiv (2) und idempotent (3).

Beweis. (1) aus $X_1 \subseteq X_2$ folgt $X_1' \supseteq X_2' \Rightarrow X_1'' \subseteq X_2'' = \varphi(X_1) \subseteq \varphi(X_2)$.
(2) ergibt sich unmittelbar aus Eigenschaft 2 des Hilfssatzes.
(3) Es gilt: $\varphi(X) = X'' = (X')' = (X')''' = X'''' = \varphi\varphi X$. □

Für jede Menge $A \subseteq G$ ist A' ein Begriffsinhalt, denn (A'', A') ist stets Begriff mit $A' = B$. Dabei ist A'' der kleinste Begriffsumfang, der A umfasst. Folglich ist eine Menge $A \subseteq G$ genau dann Begriffsumfang, wenn $A = A''$. Entsprechend ist für jede Menge $B \subseteq M$ B' ein Begriffsumfang, denn (B', B'') ist stets Begriff mit $B' = A$. Eine Menge $B \subseteq M$ ist genau dann Begriffsinhalt, wenn $B = B''$.
Die Vereinigung von Begriffsumfängen ergibt im Allgemeinen nicht wieder einen Begriffsumfang. So ergibt die Vereinigung der beiden Begriffsumfänge $(Merkur, Mars)$ und $(Jupiter, Saturn)$ aus Tabelle 2.1 keinen Begriffsumfang. Dagegen ist der Durchschnitt beliebig vieler Begriffsumfänge (bzw. Begriffsinhalte) stets wieder ein Begriffsumfang (bzw. Begriffsinhalt), wie der folgende Hilfssatz zeigt.
Dabei ist die Eigenschaft des Ableitungsoperators hilfreich, die es erlaubt, die Gemeinsamkeiten einer Vereinigung von Mengen auf die Gemeinsamkeiten der einzelnen Mengen zurückzuführen.

Hilfssatz 2.3. *Ist T eine Indexmenge und ist für jedes $t \in T$ $A_t \subseteq G$ eine Menge von Gegenständen, so ist*

$$(\bigcup_{t \in T} A_t)' \quad = \quad \bigcap_{t \in T} A_t'.$$

Entsprechendes gilt für Mengen von Merkmalen.

Beweis.

$$\begin{aligned} m \in (\bigcup_{t \in T} A_t)' &\iff gIm \text{ für alle } g \in \bigcap_{t \in T} A_t \\ &\iff gIm \text{ für alle } g \in A_t \text{ für alle } t \in T \\ &\iff m \in A_t' \text{ für alle } t \in T \\ &\iff m \in \bigcap_{t \in T} A_t'. \end{aligned}$$

□

2.3.3 Begriffsverband

Jeder formale Kontext (G, M, I) enthält eine Menge $\mathcal{B}(G, M, I)$ von Begriffen. Das Bemerkenswerte ist, dass die Begriffe einen vollständigen Verband, den Begriffsverband bilden. Dazu benötigen wir vorab einige Definitionen.

Definition 2.7. Infimum, Supremum
Es sei $(M, \leq)$ eine geordnete Menge und $A \subseteq M$. Eine untere Schranke von A ist ein

Element s von M mit $s \leq a$ für alle $a \in A$. Dual wird eine obere Schranke von A erklärt. Gibt es in der Menge aller unteren Schranken von A eine größte, so nennt man diese das Infimum von A und bezeichnet sie mit inf A oder $\bigwedge A$; dual wird eine kleinste obere Schranke Supremum genannt und mit sup A oder $\bigvee A$ bezeichnet. Ist $A = \{x, y\}$, so schreibt man für inf A auch $x \wedge y$ und für sup A auch $x \vee y$.

Definition 2.8. Vollständiger Verband
Eine geordnete Menge $V := (V, \leq)$ ist ein Verband, wenn zu je zwei Elementen x und y in V stets das Supremum $x \vee y$ und das Infimum $x \wedge y$ existieren. V heißt vollständiger Verband, falls zu jeder Teilmenge $X \subseteq V$ das Supremum $\bigvee X$ und das Infimum $\bigwedge X$ existieren. Jeder vollständige Verband V hat ein größtes Element $\bigvee V$, das als Einselement bezeichnet wird; dual wird das kleinste Element als Nullelement bezeichnet.

Alle Begriffe können in den Begriffsverband eingetragen werden. Dabei entspricht jeder Begriff im Kontext einem Knoten im Liniendiagramm. Das Liniendiagramm wird häufig auch als Hasse–Diagramm bezeichnet.
Der Hauptsatz der formalen Begriffsanalyse [GW96] besagt unter anderem, dass der Begriffsverband ein vollständiger Verband ist.

Satz 2.4. Hauptsatz über Begriffsverbände
Der Begriffsverband $\underline{\mathcal{B}}(G, M, I)$ ist ein vollständiger Verband, in dem Infimum und Supremum folgendermaßen beschrieben sind:

$$\begin{aligned} \bigwedge_{t \in T} (A_t, B_t) &= (\bigcap_{t \in T} A_t, (\bigcup_{t \in T} B_t)'') \\ \bigvee_{t \in T} (A_t, B_t) &= ((\bigcup_{t \in T} A_t)'', \bigcap_{t \in T} B_t). \end{aligned}$$

Beweis. Zunächst wollen wir die Formel für das Infimum erläutern. Da $A_t = B_t'$ für jedes $t \in T$ gilt, kann

$$(\bigcap_{t \in T} A_t, (\bigcup_{t \in T} B_t)'')$$

mit Hilfssatz 2.3 umgeformt werden zu

$$((\bigcup_{t \in T} B_t)', (\bigcup_{t \in T} B_t)''),$$

ist also von der Form (X', X'') und somit sicher ein Begriff. Dass es sich nur um das Infimum der Begriffe (A_t, B_t) handeln kann, folgt sofort daraus, dass der Umfang dieses Begriffes gerade der Durchschnitt der Umfänge der (A_t, B_t) ist. Die Formel für das Supremum wird entsprechend begründet. Damit ist nachgewiesen, dass $\underline{\mathcal{B}}(G, M, I)$ ein vollständiger Verband ist. □

Am Begriffsverband ist der Kontext noch einfacher abzulesen. Zusätzlich lässt sich aus dem System aller Begriffe der Kontext mühelos rekonstruieren. Die Bestimmung aller Begriffe eines Kontextes ist daher für die formale Begriffsanalyse von zentraler Bedeutung. Allerdings gibt es noch weitere Möglichkeiten die Zusammenhänge innerhalb eines Kontextes darzustellen, nämlich in Form von logischen Implikationen, auf die wir im Folgenden eingehen wollen.

2.3.4 Merkmalsimplikationen und Pseudoinhalte

Bei umfangreichen Kontexten bietet es sich häufig an, das Begriffssystem aus der Merkmalslogik zu erschließen, also aus den Implikationen zwischen den Merkmalen. Damit meinen wir Aussagen der Art: „Jeder Gegenstand mit den Merkmalen $a, b, c, \ldots$ hat auch die Merkmale $x, y, z \ldots$". Formal ist eine Implikation zwischen Merkmalen ein Paar von Teilmengen der Merkmalsmenge M. Wir bezeichnen eine solche Implikation mit $A \to B$ $(A, B \subseteq M)$. Die Implikationen geben Aufschluss über die Struktur des Kontextes, es ist sogar möglich anhand der Implikationen den gesamten Kontext zu rekonstruieren (siehe Beispiel 2.2). Umgekehrt können die Implikationen natürlich am Begriffsverband abgelesen werden. Dieser ist jedoch bei größeren Kontexten sehr unübersichtlich. Wie der Begriffsverband, ist die Verwendung einer miminalen Menge von Implikationen eine äquivalente Beschreibung des Kontextes. Zunächst jedoch einige Definitionen:

Definition 2.9. *Für einen Gegenstand $g \in G$ schreiben wir g' für den Gegenstandsinhalt $\{m \in M \mid gIm\}$ zum Gegenstand g. Entsprechend ist $m' := \{g \in G \mid gIm\}$ der Merkmalsumfang zu Merkmal m.*

Definition 2.10. Merkmalsimplikationen
Sei $\mathbb{K} = (G, M, I)$ ein Kontext und $A, B, T \subseteq M$. Eine Teilmenge T respektiert eine Implikation $A \to B$, wenn $A \not\subseteq T$ oder $B \subseteq T$ ist. Die Implikation $A \to B$ gilt in einem Kontext, wenn sie im System der Gegenstandsinhalte gilt. Wir sagen dann auch, $A \to B$ sei eine Implikation des Kontextes (G, M, I). Dabei bezeichnen wir A als Prämisse und B als Konklusion.

Hilfreich zur Ermittlung der Implikationen sind die folgenden Hilfssätze:

Hilfssatz 2.5. *Eine Implikation $A \to B$ gilt in (G, M, I) genau dann, wenn $B \subseteq A''$ ist. Sie gilt dann auch automatisch für die Menge aller Begriffsinhalte.*

Beweis. $B \subseteq A'' \Leftrightarrow A' \subseteq B' \Leftrightarrow [\forall g \in G : (\forall a \in A : gIa) \Rightarrow (\forall b \in B : gIb)]$.
Für Begriffsinhalte gilt ja gerade $B = B'' \Rightarrow B \subseteq B''$. □

Hilfssatz 2.6. *Ist $\mathcal{L}$ eine Menge von Implikationen in M, so ist*

$$\mathcal{H}(\mathcal{L}) := \{X \subseteq M | X \text{ respektiert } \mathcal{L}\}$$

ein Hüllensystem auf M. Ist $\mathcal{L}$ die Menge aller Implikationen eines Kontextes, dann ist $\mathcal{H}(\mathcal{L})$ das System aller Begriffsinhalte.

Beweis. $M \in \mathcal{H}(\mathcal{L})$, denn M respektiert alle Merkmalsimplikationen.
Für $\emptyset \neq \mathcal{T} \subseteq \mathcal{H}(\mathcal{L})$ ist zu zeigen, dass $\bigcap \mathcal{T}$ jede Implikation $A \to B \in \mathcal{L}$ respektiert. Dies erhalten wir wie folgt:

$$A \subseteq \bigcap \mathcal{T} \Rightarrow (\forall T \in \mathcal{T} : A \subseteq T) \Rightarrow (\forall T \in \mathcal{T} : B \subseteq T) \Rightarrow B \subseteq \bigcap \mathcal{T}.$$

□

Der zugehörige Hüllenoperator $X^{\mathcal{L}}$ wird folgendermaßen beschrieben. Für eine Menge $X \subseteq M$ sei

$$X^{\mathcal{L}} := X \cup \bigcup \{B | A \to B \in \mathcal{L}, A \subseteq X\}.$$

Man bildet die Mengen $X^{\mathcal{L}}, X^{\mathcal{L}\mathcal{L}}, X^{\mathcal{L}\mathcal{L}\mathcal{L}}, \ldots$ bis schließlich eine Menge $\mathcal{L}(X) := X^{\mathcal{L}\mathcal{L}\ldots\mathcal{L}}$ mit $\mathcal{L}(X)^{\mathcal{L}} = \mathcal{L}(X)$ entsteht. $\mathcal{L}(X)$ ist dann die gesuchte Hülle von X bezüglich des Hüllensystems $\mathcal{H}(\mathcal{L})$. Dieser Hüllenoperator entspricht der zweiten Ableitung, wie er zur Ermittlung der Implikationen in Hilfssatz 2.5 verwendet wird. Der zum Hüllensystem aller Begriffsinhalte eines Kontextes gehörende Hüllenoperator ist also: $\mathcal{H}(X) = X''$.
Mit Hilfe des Hüllensystems $\mathcal{H}(\mathcal{L})$ kann man sich zu jeder vorgegebenen Menge $\mathcal{L}$ von Implikationen einen Kontext verschaffen, dessen Begriffsinhalte genau die respektierenden Mengen sind. Das Beispiel 2.2 soll verdeutlichen, dass bei ausschließlicher Verwendung der Merkmalsimplikationen keine Informationen des ursprünglichen Kontextes verloren gehen. Aus diesem Grund können wir uns in Kapitel 4 auf die alleinige Ermittlung der Implikationen beschränken.

Beispiel 2.2. *Rekonstruktion des Kontextes mit vorgegebenen Implikationen $\mathcal{L}$*

Beispielkontext:

	1	2	3	4
a	X		X	X
b		X	X	X
c	X		X	
d		X		X

Gegeben Menge $\mathcal{L}$:

$\mathcal{L}_1$: $1 \to 3$
$\mathcal{L}_2$: $2 \to 4$
$\mathcal{L}_3$: $23 \to 4$
$\mathcal{L}_4$: $14 \to 3$
$\mathcal{L}_5$: $12 \to 34$

Hüllenoperator: $X^{\mathcal{L}} := X \cup \bigcup \{B | A \to B \in \mathcal{L}, A \subseteq X\}$

Gesucht: Alle Begriffsinhalte des Kontextes

Wähle: $X = \{1\} \to A = \{1\}$

$X^{\mathcal{L}} := \{1\} \cup \bigcup\{3\}$ *aus* $\mathcal{L}_1$,
$X^{\mathcal{L}} := \{1,3\} \Rightarrow X : \{1,3\}, A = 1$
$X^{\mathcal{L}\mathcal{L}} := \{1,3\} \cup \{3\}$
$X^{\mathcal{L}\mathcal{L}} := \{1,3\}$
$\Rightarrow \mathcal{L}(X)^{\mathcal{L}} = \mathcal{L}(X)$
$\Rightarrow \{1,3\}$ *ist Begriffsinhalt.*

Ebenso: $X = \{2\} \rightarrow A = \{2\}$
mit $\mathcal{L}_2 \Rightarrow \{2,4\}$ *ist Begriffsinhalt.*

$X = \{3\} \rightarrow A = \{3\}$
$\{3\}$ *ist Begriffsinhalt.*

Ebenso $X = \{4\}$, $\{4\}$ *ist Begriffsinhalt.*

$X = \{1,2\} \rightarrow A = \{3,4\}$
$\{1,2,3,4\}$ *ist Begriffsinhalt.*

$X = \{1,4\}$ *ist Begriffsinhalt.*

$X = \{2,3\} \rightarrow A = \{2,3\}$
aus $\mathcal{L}_3 \Rightarrow \{2,3,4\}$ *ist Begriffsinhalt.*

$X = \{3,4\} \rightarrow A = \{3,4\}$
$\{3,4\}$ *ist Begriffsinhalt.*

Damit wurden alle Begriffsinhalte des Kontextes rekonstruiert.

Außer $\mathcal{L}$ gelten im Kontext noch alle Implikationen, die aus $\mathcal{L}$ im Sinne der nachfolgenden Definition folgen:

Definition 2.11. *Sei* $\mathbb{K} = (G, M, I)$ *ein Kontext,* $A, B \subseteq M$.

- *Die Implikation* $A \rightarrow B$ *folgt aus der Implikationenmenge* $\mathcal{L} :=$

$$\forall T \subseteq M : T \text{ respektiert } \mathcal{L} \Rightarrow T \text{ respektiert } A \rightarrow B.$$

- *Die Implikationenmenge* $\mathcal{L}$ *heißt vollständig:=*

$$\text{Alle Merkmalsimplikationen von } \mathbb{K} \text{ folgen aus } \mathcal{L}$$

- *Die Implikationenmenge* $\mathcal{L}$ *heißt reduziert (nichtredundant):=*

$$\neg\exists\, A \rightarrow B \in \mathcal{L} : A \rightarrow B \text{ folgt aus } \mathcal{L} \setminus \{A \rightarrow B\}$$

Anschaulich bedeutet dies, dass eine Implikation aus $\mathcal{L}$ folgt, wenn sie in jedem Mengensystem gilt, in dem auch $\mathcal{L}$ gilt. Dazu müssen die Teilmengen, in denen eine Implikation gilt, untersucht werden. Allerdings finden sich viele Implikationen, die trivialerweise aus anderen folgen, oder in jedem Kontext gelten. Beispielsweise gilt $A \to B$ stets, wenn $B \subseteq A$ ist, weiterhin folgt aus $A \to B$ und $C \subseteq B$ stets $A \to C$. Ziel ist es daher, eine handliche vollständige Menge an Implikationen zu finden. Eliminiert man triviale Implikationen, so bleiben Implikationen mit echter Prämisse übrig.

Definition 2.12. *Für eine Merkmalsmenge $A \subseteq M$ eines Kontextes (G,M,I) bezeichnen wir mit*

$$A^* := A'' \setminus (A \cup \bigcup_{n \in A} (A \setminus \{n\})'')$$

die Menge derjenigen Merkmale, die zwar in A'', nicht aber in A oder in der Hülle einer echten Teilmenge von A liegen. Wir nennen A eine echte Prämisse, wenn $A^* \neq \emptyset$, d.h., wenn

$$A'' \neq A \cup \bigcup_{n \in A} (A \setminus \{n\})''.$$

Die Menge der Implikationen mit echter Prämisse eines Kontextes ist vollständig, wie folgender Hilfssatz zeigt:

Hilfssatz 2.7. *[GW96] Ist T eine endliche Teilmenge von M, so ist*

$$T'' \quad = \quad T \cup \bigcup \{A^* \mid A \text{ ist echte Prämisse mit } A \subseteq T\}.$$

Die Menge aller Implikationen der Form

$$A \to A^*, \quad A \text{ echte Prämisse},$$

eines Kontextes mit endlicher Merkmalsmenge ist vollständig.

Um nachzuweisen, dass eine Menge $\mathcal{L}$ von Implikationen eines Kontextes vollständig ist, muss man zeigen, dass jede Teilmenge $T \subseteq M$, die $\mathcal{L}$ respektiert, ein Begriffsinhalt ist [GW96].

Beweis. Ist $T = T''$, so ist T Begriffsinhalt, und die Behauptung trivial. Sei also $m \in T'' \setminus T$. Eine Teilmenge A von T, die minimal bezüglich der Eigenschaft $m \in A''$ ist, muss eine echte Prämisse sein, also gibt es eine Implikation $A \to A^*$ mit $m \in A^*$. Da m beliebig gewählt war, folgt die erste Behauptung. Respektiert T alle Implikationen der Form $A \to A^*$, wobei A echte Prämisse ist, so folgt aus dem soeben Bewiesenen, dass $T'' = T$ ist, also folgt, dass T ein Begriffsinhalt ist. □

Allerdings ist die so beschriebene Implikationenmenge in der Regel noch redundant. Duquenne und Guigues haben in [DG86] gezeigt, dass es zu jedem Kontext mit endlicher Merkmalsmenge M eine natürliche vollständige und nichtredundante Menge von Implikationen gibt. Diese Menge wird als Duquenne–Guigues–Basis oder auch als Stammbasis bezeichnet. Dazu wird allerdings noch der grundlegende Begriff des Pseudoinhalts benötigt, der an die Stelle der echten Prämisse tritt:

Definition 2.13. *$P \subseteq M$ heißt Pseudoinhalt von (G,M,I) genau dann, wenn $P \neq P''$ ist und für jeden Pseudoinhalt $Q \subseteq P, Q \neq P$ schon $Q'' \subseteq P$ gilt.*

Satz 2.8. *[DG86] Sei $\mathbb{K} = (G,M,I)$ ein Kontext. Dann ist die Implikationenmenge:*

$$\mathcal{L} := \{P \to P'' \mid P \quad Pseudoinhalt\}$$

vollständig und nichtredundant.

Beweis. Offenbar gilt $\mathcal{L}$ in (G,M,I). Um zu zeigen, dass $\mathcal{L}$ vollständig ist, müssen wir wieder zeigen, dass jede Menge $T \subseteq M$, die $\mathcal{L}$ respektiert, ein Begriffsinhalt ist. Jede solche Menge respektiert insbesondere alle Implikationen $Q \to Q''$, wo Q ein Pseudoinhalt und $Q \subseteq T$ ist. Angenommen $T \neq T''$, dann erfüllt T selbst die Definition des Pseudoinhaltes, und die Implikation $T \to T''$ ist in $\mathcal{L}$, wird aber nicht von T respektiert $\Rightarrow$ Widerspruch.
Um zu zeigen, dass $\mathcal{L}$ nichtredundant ist, betrachten wir einen beliebigen Pseudoinhalt P und zeigen, dass P die Menge $\mathcal{L} \setminus \{P \to P''\}$ respektiert. Ist nämlich $Q \to Q''$ eine Implikation in $\mathcal{L} \setminus \{P \to P''\}$ mit $Q \subseteq P$, dann muss auch $Q'' \subseteq P$ sein, da P ein Pseudoinhalt ist. □

Die ermittelten Implikationen der Stammbasis entsprechen den aussagenlogischen Hornausdrücken, wie sie in Unterabschnitt 2.2.3.2 eingeführt wurden. Dabei bezeichnen wir eine Implikation als einfache Implikation, wenn die Konklusion ein Element umfasst [Zic91]. Eine einfache Implikation entspricht genau einem aussagenlogischen Hornausdruck mit einem positiven Literal. Dementsprechend kann eine Menge einfacher Implikationen mehreren Hornausdrücken entsprechen.

Hornklausel(n)		Implikation
$A_1 \wedge \ldots \wedge A_n \to B$	$\leftrightarrow$	$\{A_1,\ldots,A_n\} \to B$
$\begin{pmatrix} A_1 \wedge \ldots \wedge A_n & \to & B_1 \\ \vdots & & \\ A_1 \wedge \ldots \wedge A_n & \to & B_m \end{pmatrix}$	$\leftrightarrow$	$\{A_1,\ldots,A_n\} \to \{B_1,\ldots,B_m\}$

Dabei seien $A_1,\ldots,A_n,B_1,\ldots,B_m \subseteq M$. Die aussagenlogischen Hornausdrücke besitzen genau ein positives Literal, sind also von der Gestalt $AH1$, wie sie auf Seite 10 vorgestellt wurde. In [Zic91, Gan99] wurde zudem gezeigt, dass diese Implikationen auch in die Prädikatenlogik übertragbar sind.

2.4 Statistische Grundlagen und Hypothesentests

2.4.1 Empirische Lageparameter

Zur kompakten Beschreibung empirischer Daten werden häufig sogenannte Lageparameter verwendet. Dadurch lässt sich Datenmaterial durch Angabe eines Parameters charakterisieren [BB96]. Im Folgenden wollen wir drei Lageparameter in Form von Mittelwerten kurz vorstellen [LMGR00].

- Arithmetisches Mittel:
  $\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Geometrisches Mittel:
  $x_g = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots x_n}$
- Median:
  $$x_{med} = \begin{cases} x_{(\frac{n+1}{2})} & : \text{ falls n ungerade} \\ x_{(\frac{n}{2})} & : \text{ falls n gerade} \end{cases}$$

Der Median ist durch seine Eigenschaft definiert, dass er die Datenreihe teilt. Liegen die Messwerte bereits geordnet vor $(x_1 \leq x_2 \leq \ldots \leq x_n)$, so liegt der Median in der „Mitte" der beobachteten Werte.

2.4.2 Grundbegriffe der Wahrscheinlichkeitstheorie

Ein Vorgang den wir beliebig oft wiederholen können, und der verschiedene, sich gegenseitig ausschließende Ausgänge besitzt, bezeichnen wir als Zufallsexperiment. Allgemein wird einem Zufallsexperiment mit endlich vielen Ausgängen, eine endliche nichtleere Menge Ω zugeordnet, deren Elemente ω die Versuchsausgänge bezeichnen. Mit Ω bezeichnen wir die Ergebnismenge oder den Ergebnisraum, dementsprechend nennen wir ω die Ergebnisse oder Elementarereignisse. Also gilt für Ω:

$$\Omega = \{\omega : \omega \text{ ist Elementarereignis}\}$$

Zusätzlich bezeichnen wir Teilmengen von Ω als Ereignisse. Diese werden häufig mit A oder B bezeichnet. Im Gegensatz zu den Elementarereignissen schließen sich Ereignisse nicht notwendigerweise gegenseitig aus. Die Gesamtheit aller Ereignisse bildet ein Ereignissystem, ist also eine Menge von Teilmengen von Ω, und entspricht mengentheoretisch der Potenzmenge $\mathcal{P}(\Omega)$. Bei Durchführung eines Zufallsexperimentes wollen wir jedem Ereignis $A \in \Omega$ eine Maßzahl für die Chance des Eintretens von A zuordnen, die wir ein Wahrscheinlichkeitsmaß nennen, wenn sie folgende Bedingungen erfüllt [LW00, Bau91]:

Definition 2.14. *Eine Abbildung P von $\mathcal{P}(\Omega)$ in [0,1] heißt Wahrscheinlichkeitsmaß wenn gilt:*

(i) $P(\Omega) = 1$

(ii) $P(A) \geq 0$ *für alle A*

(iii) $P(A \cup B) = P(A) + P(B)$ *für alle disjunkten A,B*

Diese Axiome werden als Axiome von Kolmogoroff oder Axiome der Wahrscheinlichkeitsrechnung bezeichnet. Das Paar (Ω, P) heißt der dem Experiment zugeordnete Wahrscheinlichkeitsraum [Kre03]. Aus diesen Eigenschaften können noch weitere hilfreiche Rechenregeln abgeleitet werden [LW00].
Bei endlicher Ergebnismenge $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ und gleichen Wahrscheinlichkeiten aller $\omega \in \Omega$, spricht man von einem Laplaceschen Wahrscheinlichkeitsraum[1]. Hier gilt für die Wahrscheinlichkeit des Ereignisses A die folgende einprägsame Regel:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{Anzahl der für A günstigen Ergebnisse}}{\text{Anzahl aller möglichen Ergebnisse}}$$

Die Ermittlung der Wahrscheinlichkeit eines Ereignisses A reduziert sich also auf empirisches Abzählen der günstigen Ausgänge für A.

2.4.3 Bedingte Wahrscheinlichkeiten

Häufig interessiert man sich für die Wahrscheinlichkeit eines Ereignisses A unter der Annahme, dass ein bestimmtes Ereignis B eintritt, oder bereits eingetreten ist. Diese bedingte Wahrscheinlichkeit ist folgendermaßen definiert:

Definition 2.15. *Sei (Ω, P) ein Wahrscheinlichkeitsraum. Sind A, B Ereignisse und gilt $P(B) > 0$, so heißt*

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

die bedingte Wahrscheinlichkeit von A unter der Bedingung B.

In der Anwendung wird jedoch häufig $P(A \cap B)$ aus $P(B)$ und $P(A|B)$ ermittelt. Dazu muss die Gleichung in Definition 2.15 lediglich in $P(A \cap B) = P(A|B) \cdot P(B)$ umformuliert werden.
Allgemein gilt für das Rechnen mit bedingten Wahrscheinlichkeiten der Satz der totalen Wahrscheinlichkeit [LW00]:

[1] Benannt nach dem französischen Mathematiker Pierre-Simon Laplace.

Satz 2.9. *Seien $B_1, B_2, \ldots, B_n$ paarweise disjunkte Ereignisse mit $\bigcup_{i=1}^{n} B_i = \Omega$ und $P(B_i) > 0$, $i = 1, \ldots, n$. Dann gilt für ein Ereignis A:*

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i) \tag{2.1}$$

Beweis. $A \in \Omega$, daher ist $A \cap \Omega = A$. Mit $\Omega = \bigcup_{i=1}^{n} B_i = \sum_{i=1}^{n} B_i$ ist $A = A \cap \sum_{i=1}^{n} B_i$ und die gesuchte Gleichheit ergibt unter Anwendung der Definition 2.15 als

$$P(A) = P(A \cap \sum_{i=1}^{n} B_i) = P(\sum_{i=1}^{n} A \cap B_i) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i). \qquad \square$$

Unter den Voraussetzungen von Satz 2.9 gilt im Falle von $P(A) > 0$ die Formel von Bayes:

$$P(B_i|A) = \frac{P(A|B_i) \cdot P(B_i)}{P(A)} \quad \text{für } i = 1, \ldots, n. \tag{2.2}$$

2.4.4 Grundbegriffe der Testtheorie

In der schließenden Statistik sollen anhand von Beobachtungen Rückschlüsse auf bestimmte Gesetzmäßigkeiten gezogen werden. Häufig sind diese Beobachtungen zahlreichen Zufallseinflüssen unterworfen, die damit keine absolut gültigen Schlüsse erlauben. Die statistische Testtheorie kann solche Fehler nicht verhindern, ist aber bestrebt die Fehler kontrollierbar bzw. möglichst kalkulierbar zu machen.
Ein Test ist ein Verfahren zur Überprüfung von Annahmen über bestimmte Verteilungen, die das Zustandekommen von Beobachtungsdaten beschreiben. Die Annahmen werden als Hypothesen bezeichnet. Ein Test ist also eine Entscheidungsregel, die für jeden beobachteten Wert festlegt, ob man sich für oder gegen die Hypothese entscheidet. Im ersten Fall wird die Hypothese (Nullhypothese) angenommen, der zweite Fall führt zum Verwerfen der Nullhypothese. Mit der Formulierung einer Nullhypothese (H_0) wird auch stets eine Alternativhypothese (H_1) angegeben. Sie stellt den gegenteiligen Wert der Nullhypothese dar, und wird angenommen, wenn H_0 verworfen wird. Ein Test ist durch die Angabe eines Verwerfungsbereiches bzw. kritischen Bereiches beschrieben. Soll eine Entscheidung bezüglich H_0 getroffen werden, muss stets ein kritischer Wert angegeben werden, der über Annahme oder Verwerfen entscheidet.
Durch Annahme oder Verwerfen von Hypothesen können auch Fehlentscheidungen getroffen werden. Generell sind zwei Arten von Fehlern möglich. Ist H_0 richtig und wird fälschlicherweise trotzdem verworfen, so spricht man von einem Fehler der ersten Art. Die Wahrscheinlichkeit H_0 fälschlicherweise zu verwerfen wird mit α bezeichnet. Von einem Fehler der zweiten Art spricht man, wenn H_0 zu Unrecht angenommen wird, die Wahrscheinlichkeit dafür wird mit β bezeichnet.

Die Wahrscheinlichkeit einen Fehler der ersten Art zu begehen wird auch als Signifikanzniveau α eines Tests bezeichnet. Das Signifikanzniveau wird in der Regel vor der Durchführung des Tests gewählt, so dass der Fehler der ersten Art direkt kontrolliert werden kann. Mit Wahl von α ist auch der kritische Bereich festgelegt. Da die Anzahl der Beobachtungen stark unter oder auch stark über dem kritischen Bereich liegen kann, existieren zwei Verwerfungsbereiche. In diesem Fall wird jeweils $\alpha/2$ für den oberen und unteren Grenzbereich festgelegt.

2.4.5 Hypothesentests

Im folgenden Abschnitt werden statistische Hypothesentests vorgestellt, die wir zu einem späteren Zeitpunkt in dieser Arbeit verwenden wollen. In vielen Testverfahren wird eine Normalverteilung des zugrunde liegenden Sachverhaltes angenommen. Wir wollen allerdings auf Testverfahren eingehen, die bei beliebigen Verteilungsannahmen anwendbar sind, also verteilungsfrei sind. Dabei wird eine Teststatistik verwendet, die durch eine χ^2–Verteilung approximiert werden kann.

- **χ^2–Unabhängigkeitstest**
  Beim χ^2–Unabhängigkeitstest soll für eine zweidimensionale Zufallsvariable (X,Y) überprüft werden, ob ihre Komponenten X und Y stochastisch unabhängig sind. Dazu wird der Wertebereich der Komponenten in disjunkte Intervalle zerlegt. Sei $A_1,\ldots,A_k$ eine disjunkte Zerlegung von X, und $B_1,\ldots,B_l$ die Zerlegung von Y. Für $i=1,\ldots,k$ und $j=1,\ldots,l$ gilt:

$$\begin{aligned} p_{ij} &= P(X \in A_i, Y \in B_j) \quad \text{und} \\ p_{i\bullet} &= \sum_{j=1}^{l} p_{ij} = P(X \in A_i) \quad \text{sowie} \\ p_{\bullet j} &= \sum_{i=1}^{k} p_{ij} = P(X \in B_j) \end{aligned}$$

  Falls X und Y stochastisch unabhängig sind müsste gelten:

$$p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$$

  Weiterhin bezeichnen wir für Stichproben $(X_1,Y_1),\ldots,(X_n,Y_n)$:

$$\begin{aligned} N_{ij} &= \text{Anzahl der } m \in \{1,\ldots,n\} \text{ mit } X_m \in A_i \text{ und } Y_m \in B_j \\ N_{i\bullet} &= \sum_{j=1}^{l} N_{ij} = \text{Anzahl der } m \text{ mit } X_m \in A_i \\ N_{\bullet j} &= \sum_{i=1}^{k} N_{ij} = \text{Anzahl der } m \text{ mit } Y_m \in B_j \end{aligned}$$

Es wird die Nullhypothese

$$H_0 : p_{ij} \quad = \quad p_{i\bullet} \cdot p_{\bullet j} \quad \text{für alle Paare } (i,j)$$

zur Alternativhypothese

$$H_1 : p_{ij} \quad \neq \quad p_{i\bullet} \cdot p_{\bullet j} \quad \text{für mindestens ein Paar } (i,j)$$

überprüft. Als Testgröße wird die quadratische Abweichung der beobachteten Häufigkeiten ermittelt.

$$T \quad = \quad \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(N_{ij} - n \cdot p_{i\bullet} \cdot p_{\bullet j})^2}{n \cdot p_{i\bullet} \cdot p_{\bullet j}}$$

Allerdings sind $p_{i\bullet}$ und $p_{\bullet j}$ nicht bekannt und werden daher über ihre relativen Häufigkeiten $\hat{p}_{i\bullet} = \frac{N_{k\bullet}}{n}$ sowie $\hat{p}_{\bullet j} = \frac{N_{\bullet l}}{n}$ abgeschätzt. Daraus ergibt sich:

$$T \quad = \quad \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(N_{ij} - \frac{N_{i\bullet} \cdot N_{\bullet j}}{n})^2}{\frac{N_{i\bullet} \cdot N_{\bullet j}}{n}}$$

Durch Umformung erhält man:

$$T \quad = \quad n \cdot (\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{N_{ij}^2}{N_{i\bullet} \cdot N_{\bullet j}} - 1)$$

Die Nullhypothese H_0 ist gültig, wenn die Prüfstatistik T kleine Werte annimmt. Sie ist für große n näherungsweise $\chi^2_{(k-1)(l-1)}$–verteilt [WN70]. Daher wird H_0 abgelehnt, wenn

$$T \quad = \quad n \cdot (\sum_{i=1}^{k} \sum_{j=1}^{l} \frac{N_{ij}^2}{N_{i\bullet} \cdot N_{\bullet j}} - 1) > \chi^2_{(k-1)(l-1);1-\alpha}$$

ist. Die Größe $\chi^2_{(k-1)(l-1);1-\alpha}$ wird auch als das $(1-\alpha)$–Quantil der χ^2–Verteilung bezeichnet.

- **Exakter Test von Fisher**
  Bei Vorlage einer Vierfeldertafel wird häufig zur Überprüfung der Unabhängigkeit zweier Merkmale der exakte Test von Fisher verwendet. Dieser bietet sich vor allem für kleinere Stichproben an. Der Vorteil des Tests liegt darin, dass die Testgrößen nicht nur näherungsweise, sondern exakt angegeben werden können. Um eine quantitative Aussage über das Auftreten der Merkmale machen zu können, werden aus einer dichotomen Menge Stichproben entnommen und auf eine bestimmte Eigenschaft hin untersucht. In der Statistik wird dies als „Ziehen ohne Zurücklegen" formuliert. Die Zufallsvariable, die diesen Vorgang beschreibt, ist eine hypergeometrisch verteilte Zufallsvariable gemäß der folgenden Definition [LW00].

Definition 2.16. *Seien $n,N,M \in \mathbb{N}$, und gelte $n,M \leq N$. Die Zufallsvariable X heißt hypergeometrisch verteilt mit Parametern n,N,M (kurz ($H(n,N,M)-verteilt$), falls*

$$P(X=i) \quad = \quad \frac{\binom{M}{i} \cdot \binom{N-M}{n-i}}{\binom{N}{n}}, \qquad i = max(0,n-N+M),\ldots,min(n,M)$$

gilt.

Damit kann das Vorliegen bestimmter Merkmale in der Stichprobe, durch eine hypergeometrisch verteilte Zufallsvariable X beschrieben werden. Nach Wahl des Signifikanzniveaus α ermittelt man die kritischen Werte der Verteilung. Die Quantile der hypergeometrischen Verteilung können z. B. in statistischen Tafelwerken nachgeschlagen werden.

Eine ausführliche Darstellung weiterer Hypothesentests findet man z. B. in [BLK06, LW00, Kre03].

2.5 Auswertung von Klassifikationsergebnissen

2.5.1 ROC–Graphen

ROC–Graphen[2] sind eine Möglichkeit Klassifikationsergebnisse auszuwerten und zu visualisieren. Sie werden intensiv im medizinischen Bereich zur Bewertung von Behandlungsstrategien verwendet. Grundlage für die ROC–Analyse ist die Kontingenztabelle. Wir gehen im Folgenden stets von einem Klassifikationsproblem mit zwei Klassen (z. B. Kreditausfall/kein Kreditausfall oder krank/gesund) aus. Diese Klassen bezeichnen wir mit $\{p,n\}$. Dabei bezeichnet p die positiven Klassen (z. B. krank), n bezeichnet die negativen Klassen (z. B. gesund). Ein Klassifikationsmodell ordnet die Datensätze vorhergesagten diskreten Klassen zu. Um die vorhergesagten Klassen von den Ursprungsklassen zu unterscheiden bezeichnen wir diese mit $\{Y,N\}$. Eine vorhergesagte positive Klasse nennen wir Y, eine vorhergesagte negative Klasse bezeichnen wir mit N.
Sind ein Klassifikationsmodell und ein Datensatz gegeben, so existieren vier mögliche Klassifikationsentscheidungen. Ist der Datensatz positiv und das Klassifikationsergebnis ebenfalls, so bezeichnen wir dies als True Positive, wird er negativ eingestuft, so wird er False Negative genannt. Handelt es sich um einen negativen Datensatz, kann dieser einer negativen oder positiven Klasse zugeordnet werden. Den ersten Fall kennzeichnen wir mit True Negative, den zweiten Fall mit False Positive. Die vier möglichen Zuordnungen sind in Tabelle 2.2 dargestellt. Die Einträge entlang der Hauptdia-

[2] ROC = Receiver–Operating–Characteristic

	p	n
Y	**T**rue **P**ositives	**F**alse **P**ositives
N	**F**alse **N**egatives	**T**rue **N**egatives
	$\sum = P$	$\sum = N$

Tabelle 2.2: Kontingenztabelle mit möglichen Klassifikationsergebnissen

gonalen stellen die Anzahl der richtig klassifizierten Fälle dar. Die Summe der ersten Spalte beziffert alle echt positiven Fälle P, die Summe der zweiten Spalte alle echt negativen Fälle N. Folgende Kenngrößen lassen sich aus der Kontingenztabelle ableiten [Faw03].

$$\text{false positive rate} = \frac{FP}{N} \qquad \text{precision} = \frac{TP}{TP+FP}$$

$$\text{true positive rate} = \frac{TP}{P} \qquad \text{accuracy} = \frac{TP+TN}{P+N}$$

Dabei kennzeichnet die true positive rate den Anteil aller richtig positiv klassifizierten Fälle, bezogen auf die Summe aller echt positiven Fälle. Die false positive rate kennzeichnet den Anteil aller fälschlicherweise positiv klassifizierten Fälle, bezogen auf die Summe aller echt negativen Fälle. Die accuracy wird häufig auch als Gesamtklassifikationsgüte bezeichnet. Zur Ermittlung der ROC–Kurven werden Sensitivität und Spezifität benötigt. Sie liegen stets im Intervall $[0,1]$ und sind folgendermaßen definiert:

$$\text{Sensitivität} = \frac{TP}{P} = \frac{TP}{TP+FN} \tag{2.3}$$

$$\text{Spezifität} = 1 - \text{false positiv rate} = \frac{TN}{TN+FP} \tag{2.4}$$

ROC–Kurven werden in zweidimensionalen Graphen dargestellt. Üblicherweise wird auf der y–Achse die true positive rate (Sensitivität) und auf der x–Achse die false positive rate (1 – Spezifität) abgetragen. Abbildung 2.1 zeigt beispielhaft einen ROC–Graphen. Im Punkt $(0,0)$ des ROC–Graphen wurde kein positiver Datensatz richtig klassifiziert, allerdings wurde auch kein Datensatz fälschlicherweise positiv klassifiziert. Ein gegensätzliches Ergebnis liefert der Punkt $(1,1)$. Hier werden alle Datensätze positiv bewertet. Ein perfektes Klassifikationsergebnis wird im Punkt $(0,1)$ erreicht. Daher gilt generell in ROC–Graphen, dass ein Punkt (x,y) „besser" einzustufen ist, je näher er am „perfekten" Punkt $(0,1)$ liegt. Daneben ist in Abbildung 2.1 die Diagonale $y = x$ eingezeichnet, die das Ergebnis des „bloßen Ratens" repräsentiert. Ein realistisches Klassifikationsergebnis sollte daher stets über der Diagonalen liegen.
Allerdings existieren diskrete Klassifikationsmodelle, wie z. B. Entscheidungsbäume, die für jeden Datensatz genau ein Ergebnis liefern. Angewendet auf eine Menge von Testdaten produziert das Verfahren eine Kontingenztabelle und liefert daher nur einen

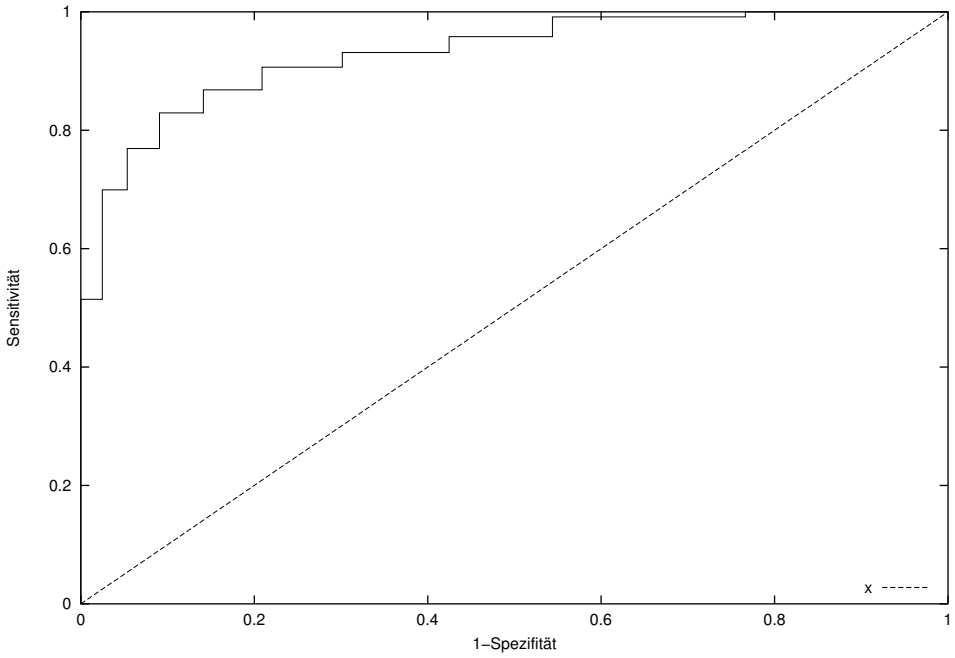


Abbildung 2.1: Beispielhafte Darstellung eines ROC–Graphen

Wert für Sensitivität und Spezifität, was einem Punkt im ROC–Graphen entspricht. In [PD03] werden Verfahren beschrieben, die es möglich machen, aus diskreten Ergebnissen ROC–Graphen zu generieren.

2.5.2 AUC–Werte

Um verschiedene ROC–Graphen miteinander zu vergleichen, ist man an einem quantitativen Maß interessiert. Ein solches Maß stellt dabei die Fläche unter dem ROC–Graphen dar. Sie wird als AUC–Wert[3] bezeichnet und liegt stets im Intervall $[0,1]$. Die in Abbildung 2.1 dargestellte Diagonale $y = x$ halbiert das Einheitsquadrat und besitzt daher einen AUC–Wert von 0.5. Die Fläche unter der ROC–Kurve hat eine anschauliche Bedeutung. Sie gibt die Wahrscheinlichkeit an, dass ein zufällig ausgewählter positiver Datensatz höher bewertet wird, als ein zufällig ausgewählter negativer Datensatz. In [HM82] wurde gezeigt, dass die Ermittlung des AUC–Wertes dem Wilcoxon–Rang–Test bzw. der Mann–Whitney–Statistik entspricht. Der AUC–Wert einer ROC–Kurve kann leicht mittels Trapezoid–Methode berechnet werden. Dazu wird der Graph mit Hilfe der Datenpunkte in Trapeze zerlegt und ein linearer Verlauf des Graphen zwischen den Datenpunkten angenommen.
Beim Vergleich zweier ROC–Graphen anhand des AUC–Wertes sollte allerdings die

[3] AUC = Area Under the (ROC) Curve

Gestalt beider Kurven berücksichtigt werden. Es ist z. B. bei sich schneidenen ROC–Kurven möglich, dass eine Kurve einen höheren AUC–Wert als der Vergleichsgraph aufweist, der jedoch in bestimmten Intervallen ein besseres Klassifikationsergebnis erzielt. Eine ROC–Analyse sollte daher den AUC–Wert und die Gestalt der Kurve umfassen.

2.6 Bauspartechnische Grundlagen

Bausparen ist ein freiwilliges Sparen mit dem Ziel ein Bauspardarlehen für wohnwirtschaftliche Zwecke zu erlangen. Die Verzinsung des Bausparguthabens und des Bauspardarlehens sind mit Abschluss des Bausparvertrages über eine bestimmte Bausparsumme festgelegt, und damit von Zinsschwankungen am Kapitalmarkt unabhängig. Ein Bausparvertrag gliedert sich in vier Phasen [Lau05]:

1. Abschluss- bzw. Einlösephase
2. Sparphase
3. Zuteilung bzw. Auszahlungsphase
4. Darlehens- bzw. Tilgungsphase

Wird die Abschlussgebühr direkt bei Abschluss des Bausparvertrages geleistet, so befindet sich der Bausparer sofort in der Sparphase. Während der Sparphase zahlt der Bausparer auf sein Konto ein, und stellt damit die Bauspareinlagen dem Bausparkollektiv zur Verfügung. Während der Sparphase ist der Bausparer in seinem Sparverhalten völlig frei. Die Sparphase endet mit der Zuteilung des Bausparvertrages und ist von der Sparleistung des Sparers abhängig. Mit der Zuteilung tritt der Bausparer auch in die letzte Phase, die Darlehensphase ein. Dabei kann der Bausparer von der Bausparkasse ein Darlehen verlangen, dessen maximale Höhe sich aus der Differenz der vereinbarten Bausparsumme zum Bausparguthaben errechnet [TCDL02]. Der Bausparer ist im Gegenzug verpflichtet, das Bauspardarlehen mit einer von der Bausparkasse vorgeschriebenen Mindesttilgung zu tilgen.
Abbildung 2.2 zeigt einen idealisierten Ablauf eines Bausparkontos. In der Realität kann der Ablauf des Bausparvertrages allerdings stark von dem dargestellten Ablauf abweichen, da dem Bausparer innerhalb der verschiedenen Phasen unterschiedliche Optionen zur Verfügung stehen.

- **Kündigung**
  Der Bausparer kann seinen Bausparvertrag innerhalb der Sparphase jederzeit kündigen. Hat der Bausparer staatliche Wohnungsbauprämie erhalten, so unterscheidet man zwischen Kündigung innerhalb und außerhalb der Sperrfrist, die derzeit 7 Jahre beträgt. Kündigt ein Bausparer innerhalb der Sperrfrist, so muss

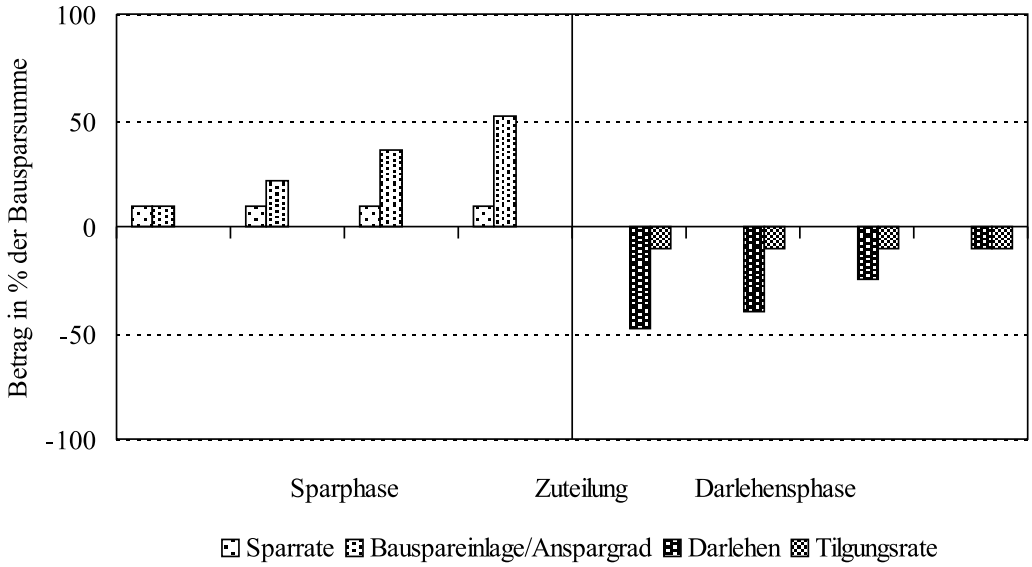


Abbildung 2.2: Idealisierter Kontoverlauf mit konstanter Sparzahlung und konstanter Tilgung. Der Anspargrad ist der Anteil der Bauspareinlage an der Bausparsumme

er die erhaltene Wohnungsbauprämie zurückzahlen. Wurde keine Wohnungsbauprämie vom Bausparer bezogen, so ist es unerheblich, ob die Kündigung innerhalb oder außerhalb der Sperrfrist stattfindet.

- **Fortsetzung**
  Wenn der Bausparer die Zuteilungsvoraussetzung erfüllt und die ihm angebotene Zuteilung nicht annimmt bzw. keine Absicht der Zuteilungsannahme äußert, so wird der Bausparvertrag fortgesetzt [Lau05]. Zu einem späteren Zeitpunkt kann der Bausparer sein Recht auf Zuteilung wiedergeltend machen.

- **Auszahlungsverschübe**
  Mit der Zuteilungsannahme stellt die Bausparkasse die Bausparsumme zur Auszahlung bereit. Allerdings kann der Bausparer sich das Bausparguthaben und das Bauspardarlehen zu verschiedenen Zeitpunkten auszahlen lassen. Nach Auszahlung des Guthabens können durchaus Jahre vergehen bis der Bausparer das Darlehen abruft. Allerdings kann die Bausparkasse – wie andere Kreditinstitute auch – Bereitstellungszinsen vom Bausparer verlangen [Lau05].

- **Sondertilgung**
  Der Bausparer ist berechtigt, neben der von der Bausparkasse vorgeschriebenen tariflichen Mindesttilgung, jederzeit Sondertilgungen zu leisten. Dafür ist vom Bausparer keine Vorfälligkeitsentschädigung zu entrichten. Sondertilgungen finden häufig in Form von Ablösungen des gesamten Darlehens statt.

- **Darlehensverzicht**
  Nach der Zuteilung ist der Bausparer nicht verpflichtet sein Darlehen in Anspruch zu nehmen. Er kann ganz oder teilweise auf seinen Darlehensanspruch verzichten. Dies ist bei hoch verzinsten Bausparverträgen häufig der Fall, da der Bausparer am Kapitalmarkt eventuell bessere Konditionen vorfindet.

Daneben existieren weitere Formen des Bausparens, die sogenannten Vor- und Zwischenfinanzierungsverträge (VK/ZK–Verträge). Beim Vorfinanzierungsvertrag gewährt die Bausparkasse dem Bausparer ein Vorausdarlehen. Im Gegenzug wird vom Darlehensempfänger ein Bausparvertrag über die volle Summe des Vorausdarlehens abgeschlossen. Allerdings ist der Bausparer bei der Besparung dieses Vertrages nicht frei, sondern muss sich an die vorgeschriebenen Sparraten der Bausparkasse halten.
Beim Zwischenfinanzierungsvertrag ist beispielsweise ein bestehender Bausparvertrag noch nicht zuteilungsreif. Die Bausparkasse gewährt einen Zwischenkredit über die volle Bausparsumme des nicht zuteilungsreifen Bausparvertrages. Mit Zuteilung des Bausparvertrages wird der Zwischenkredit getilgt. Nähere Details zu Vor- und Zwischenfinanzierungsverträgen finden sich z. B. unter [Lau05, TCDL02].

2.7 Kreditausfallwahrscheinlichkeiten

2.7.1 Problemstellung

In einem ersten Schritt soll der zentrale Begriff des Kreditrisikos erläutert werden. Unter Kreditrisiko verstehen wir im Rahmen dieser Arbeit das Risiko, dass ein Schuldner seinen Verpflichtungen bestehend aus Zins- und Tilgungszahlungen innerhalb eines bestimmten Zeitraumes nicht nachkommt. Dieser Zeitraum wird in der Rahmenvereinbarung des Baseler Ausschusses für Bankenaufsicht mit 90 Tagen angegeben [Bas04]. Kreditausfälle liegen in der Zukunft und basieren auf Prognosen. Daher liegt das Ziel dieser Arbeit unter anderem darin, Prognosen für die Kreditausfälle in Form von Kreditausfallwahrscheinlichkeiten zu ermitteln. In der vorliegenden Arbeit werden schuldnerspezifische Ausfallwahrscheinlichkeiten von Privatkrediten untersucht. Dabei gewinnt die Betrachtung der Kreditausfallwahrscheinlichkeiten von Privatpersonen zunehmend an Bedeutung. Die Anzahl der privaten Verbraucherinsolvenzen steigt im Gegensatz zu den Unternehmensinsolvenzen kontinuierlich an (siehe Abbildung 2.3). Sie umfasste im 1. Halbjahr 2006 bereits 43.600 Personen [Cre06]. Unter diesem Aspekt, und auch im Hinblick auf die Umsetzung von Basel II, erscheint eine risikogerechte Ermittlung der Ausfallwahrscheinlichkeiten notwendig. Die genaueren Verfahren zur Risikomessung führen zudem zu Erleichterungen bei der notwendigen Eigenkapitalausstattung der Kreditinstitute.

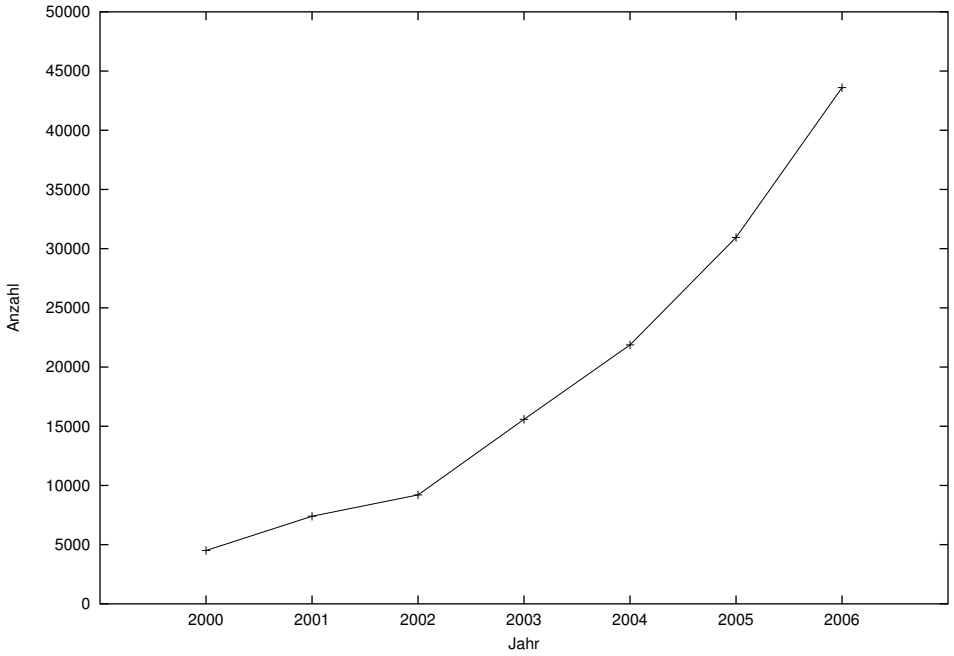


Abbildung 2.3: Entwicklung der Verbraucherinsolvenzen seit dem Jahr 2000 (jeweils 1. Halbjahr) [Cre06].

2.7.2 Quantifizierung des Kreditrisikos

Neben den schuldnerspezifischen Ausfallwahrscheinlichkeiten sind in der Praxis weitere Risikomaße von Interesse, die dazu beitragen das Kreditrisiko zu quantifizieren. Die den Verlust bestimmenden Risikoparameter sind im einzelnen [BCK04]:

- **PD (Kreditausfallwahrscheinlichkeit)**
  Mit PD wird die Wahrscheinlichkeit bezeichnet, dass ein Kreditnehmer zahlungsunfähig wird (PD = Probability of Default). Sie ist abhängig von der Bonität des Kreditnehmers.

- **EAD (Kreditbetrag bei Ausfall)**
  Der Risikoparameter EAD gibt an, wie hoch die Kreditbenutzung zum Zeitpunkt des Kreditausfalls sein wird. Er entspricht dem erwarteten Forderungsbetrag zum Ausfallzeitpunkt (EAD = Exposure at Default).

- **LGD (Verlustquote bei Ausfall)**
  Die Verlustquote gibt an, welchen Teil der Forderung die Bank bei Ausfall des Kredits verlieren wird (LGD = Loss Given Default). Die Verlustquote hängt von Art und Wert der Sicherheiten und ihrer Liquidierbarkeit ab.

- **M (Restlaufzeit)**
  Der Risikoparameter M bezeichnet die effektive (Rest)–Laufzeit des Kredites (M = Maturity). Diese wird benötigt, da die Tendenz besteht, dass Kreditnehmer im Zeitraum der Kreditlaufzeit ihre Bonität verändern [Neu98].

Diese Parameter ermöglichen nun eine Quantifizierung des Kreditrisikos als Kostenfaktor. Dabei wird generell zwischen erwarteten und unerwarteten Verlusten unterschieden. Der erwartete Verlust (EL = Expected Loss) versucht die dem Kreditgeschäft inhärenten Verlustrisiken abzubilden und ermittelt sich aus dem folgenden Produkt [Deu04]:

$$EL = PD \cdot EAD \cdot LGD. \tag{2.5}$$

Die erwarteten Verluste sind eine kalkulierbare Kostenkomponente im Kreditgeschäft und werden von den Instituten bereits durch Wertberichtigungen und Zinsmargen abgedeckt. Das tatsächliche Kreditrisiko besteht allerdings aus der Ungewissheit, ob sich der erwartete Verlust tatsächlich einstellt, oder ob die effektiven Verluste höher ausfallen. Die Abweichungen vom erwarteten Verlust werden als unerwarteter Verlust (UL = Unexpected Loss) bezeichnet. Der unerwartete Verlust wird in [Deu04] folgendermaßen quantifiziert:

$$UL = EAD \cdot LGD - EL \tag{2.6}$$

Die tatsächlichen Abweichungen vom EL können sehr hoch sein, treten dafür aber relativ selten auf. Ein typisches Beispiel dafür ist der gleichzeitige Ausfall vieler Kreditnehmer aufgrund einer Rezession.
Im folgenden Abschnitt werden die Vorschläge des Baseler Ausschuss für Bankenaufsicht zur Eigenkapitalhinterlegung vorgestellt, die sich bei der Kalibrierung der Risikogewichte am Konzept der unerwarteten Verluste orientierten [WBE04, WEV01].

2.7.3 Neue Eigenkapitalanforderungen für Kreditinstitute (Basel II)

Wesentliches Ziel der neuen Eigenkapitalrichtlinien Basel II ist es, die Kapitalanforderungen stärker als bisher vom eingegangenen Risiko abhängig zu machen. Die Eigenkapitalunterlegung der Kredite soll an der Bonität der Kreditnehmer ausgerichtet werden. Zudem bietet Basel II die Möglichkeit, Sicherheiten in die Risikoermittlung miteinzubeziehen. Ein zentraler Punkt der neuen Vereinbarung sind die quantitativen Eigenkapitalanforderungen, die aus unterschiedlich komplexen Verfahren zur Risikomessung resultieren.
Dabei unterscheidet Basel II zwischen den Standard– und IRB–Ansätzen. Die Standardansätze gliedern sich in den einfachen und umfassenden Ansatz. Sie basieren auf externen Ratings und unterscheiden sich nur in Bezug auf Behandlung und den Kreis

zulässiger Sicherheiten [CDEL05]. Die Standardansätze weisen den verschiedenen Schuldnerkategorien (Ratingklassen) pauschale Risikogewichte zu. Eine beispielhafte Darstellung der erforderlichen Eigenkapitalunterlegung findet sich in Tabelle 2.3. Grundsätzlich gilt: je höher das Risikogewicht, desto höher das zu hinterlegende Eigenkapital.

Aufsichtsrechtliches Verfahren	**Kapitalunterlegung von unbesicherten Unternehmenskrediten**						
	Kredit in Euro		Risiko-gewicht		Unter-legungsatz		Eigenkapital-hinterlegung der Bank
Bisherige Regelung	100.000	x	100 %	x	8 %	=	8.000 Euro
Künftige Regelung im Standardansatz	100.000	x	20 %[4]	x	8 %	=	1.600 Euro
		x	50 %	x	8 %	=	4.000 Euro
		x	100 %	x	8 %	=	8.000 Euro
		x	150 %	x	8 %	=	12.000 Euro
Künftige Regelung im IRB–Ansatz	100.000[5]	x	13 %[6] (bis)	x	8 %	=	1.040 Euro
		x	245 %	x	8 %	=	19.600 Euro

Tabelle 2.3: Grundschema aufsichtsrechtlicher Kapitalanforderungen für Banken[7]

Im Rahmen der IRB–Ansätze setzen die Kreditinstitute eigene Risikobewertungsverfahren ein um die notwendige Eigenkapitalhinterlegung zu bestimmen. Die Banken beurteilen Kreditnehmer anhand interner Rating– bzw. Risikomessverfahren, wobei ein spezielles Verfahren nicht vorgeschrieben wird. Nach den Vorschlägen des Baseler Ausschusses für Bankenaufsicht müssen die Kreditinstitute die Kreditnehmer dann in eine von mindestens acht Ratingklassen einteilen [Bas04]. Zur Ermittlung der von Basel II geforderten Eigenkapitalhinterlegung werden die in Unterabschnitt 2.7.2 einge-

[4] Pauschale Risikogewichte

[5] Im IRB–Ansatz wird als Berechnungsgrundlage anstelle des Kreditbetrags die erwartete Forderungshöhe im Ausfallzeitpunkt (EAD) verwendet.

[6] Auf Basis der Risikoparameter Ausfallwahrscheinlichkeit (PD), Verlust bei Ausfall (LGD) und Restlaufzeit (M) ermittelte Risikogewichte.

[7] Quelle: `www.basel2.helaba.de`

führten Risikoparameter benötigt. Diese Eingangsparameter werden je nach verwendetem Ansatz teilweise vorgegeben (IRB–Basisansatz) oder vollständig von der Bank geschätzt (Fortgeschrittener IRB–Ansatz). Die Risikogewichte werden mit Hilfe von aufsichtlichen Risikogewichtsfunktionen ermittelt und orientieren sich an der Art der zugrunde liegenden Forderung. Die Risikogewichtsfunktionen sind in [Deu04] ausführlich dargestellt. Das geforderte regulatorische Eigenkapital soll die unerwarteten Verluste absichern. Die aufsichtlich geforderte Eigenmittelunterlegung für Forderungen an Privatkunden (RWA) im IRB–Ansatz wird aus dem Produkt

$$RWA \quad = \quad 0.08 \cdot EAD \cdot RW(PD, LGD) \tag{2.7}$$

ermittelt. Dabei wird im Basisansatz nur der Risikoparameter PD durch die Bank geschätzt, die anderen Risikokomponenten werden aufsichtlich vorgegeben, während im fortgeschrittenen Ansatz alle Komponenten vom Kreditinstitut ermittelt werden. Der stufenweise Aufbau der IRB–Ansätze soll dazu dienen, den Übergang vom Basisansatz zum fortgeschrittenen Ansatz für die Kreditinstitute zu erleichtern. Eine beispielhafte Berechnung der geforderten Eigenkapitalhinterlegung in den verschiedenen Ansätzen findet sich in Tabelle 2.3.
Zudem besteht bei allen Ansätzen die Möglichkeit, das zu unterlegende Eigenkapital durch sogenannte Methoden der Risikominderung (Anrechnung von Garantien und Sicherheiten des Kreditnehmers) weiter zu reduzieren. Dabei richtet sich die Anerkennung von Sicherheiten nach dem verwendeten Ansatz. Die Möglichkeiten der Kreditrisikominderung sind in Tabelle 2.3 nicht berücksichtigt. Eine ausführliche Darstellung der Kreditrisikominderungstechniken findet sich ebenfalls in [Deu04, CDEL05].

Kapitel 3

Vorstellung verschiedener Klassifikationsverfahren

3.1 Neuronale Netze

3.1.1 Grundidee

Ausgangspunkt der Beschäftigung mit künstlichen neuronalen Netzen ist in der Regel das menschliche Gehirn. Dort müssen Informationen gefiltert und verarbeitet werden um zu Entscheidungen zu gelangen. Ein Neuron besitzt zur Informationsverarbeitung drei wesentliche Komponenten [RMS90]:

- einen Information empfangenden Bereich,
- einen Information verarbeitenden Bereich
- und einen Information sendenden Bereich.

Diese Komponenten werden zur Modellierung mit neuronalen Netzen verwendet.

3.1.2 Modellierung

Der mathematische Formalismus eines einzelnen Neuron stellt ein Modell dieser drei Komponenten dar. In einer Neuronenzelle werden die verschiedenen Informationskanäle zusammengeführt und gebündelt. Die Informationen sind mit x_i, $i \in \{1,\ldots,n\}$ gekennzeichnet und liegen in Form von Datensätzen vor. Mit α_i, $i \in \{1,\ldots,n\}$ werden die Gewichte der Informationskanäle bezeichnet. Ist ein Gewicht null, so wird dieser Kanal abgetrennt. Als Absolutwerte können die Gewichte als Stärke der Verbindung interpretiert werden, das Vorzeichen signalisiert dabei eine hemmende ($\alpha_i < 0$) bzw. eine verstärkende ($\alpha_i > 0$) Wirkung.
Im nächsten Schritt werden in der Neuronenzelle alle eingehenden Informationen zu einer Gesamtgröße überlagert. Dazu wird die gewichtete Summe $y = \sum_{i=1}^{n} \alpha_i x_i$ aller

gefilterten Informationen gebildet. Diese gewichtete Summe wird als Nettoinput bezeichnet.
Der nächste Schritt stellt die Entscheidungsfindung dar. Anhand der bisherigen Informationen soll eine Ja–Nein–Entscheidung getroffen werden. Im einfachsten Fall vergleicht man den Nettoinput mit einem Schwellwert α_0. Ist der Nettoinput größer als der Schwellwert, so soll der Output des Neurons $F(y)$ gleich eins sein, im anderen Fall gleich null. Im mathematischen Modell lässt sich dies durch die Heaviside Sprungfunktion auf die Differenz $y - \alpha_0$ abbilden.

$$F(y) = \begin{cases} 1; & falls \quad y - \alpha_0 > 0 \\ 0; & falls \quad y - \alpha_0 \leq 0 \end{cases} \tag{3.1}$$

Diese Funktion wird im Kontext der neuronale Netze als Aktivierungsfunktion bezeichnet. Allerdings ist der Gebrauch der Sprungfunktion stark eingeschränkt, da sie nicht überall differenzierbar ist und daher in der Netzwerkoptimierung keine Verwendung findet. Daher werden in jedem Punkt des Definitionsbereiches differenzierbare nichtlineare Aktivierungsfunktionen verwendet. Diese werden als Sigmoidfunktion bezeichnet.
Abbildung 3.2 zeigt ein Netz mit mehreren Neuronen in dem diese als einzelne Schichten angeordnet sind, deren Verbindung stets von einer Schicht in die nächste führt. Dabei werden die Neuronen der Eingabeschicht als Eingabeneuronen, die der mittleren Schicht als verdeckte Neuronen und die der Ausgangsschicht als Ausgabeneuronen bezeichnet. Der Nettoinput y_j der verdeckten Neuronen wird folgendermaßen ermittelt:

$$y_j = \sum_{i=1}^{n} x_i \alpha_{ij} \tag{3.2}$$

Als Aktivierungsfunktion f_0 wird häufig die logistische Funktion gewählt, da diese in jedem Punkt ihres Definitionsbereiches differenzierbar ist. Abbildung 3.1 zeigt die logistische Aktivierungsfunktion deren Wertebereich im Intervall $[0,1]$ liegt. Das Ausgabesignal der verdeckten Neuronen wird durch die logistische Aktivierungsfunktion folgendermaßen ermittelt:

$$f(y_j) = \frac{1}{1+e^{-y_j}} \tag{3.3}$$

Die Eingangssignale für die Ausgabeneuronen z_k setzen sich daher folgendermaßen zusammen:

$$z_k = \sum_{j=1}^{m} \beta_{jk} f(y_j) \tag{3.4}$$

Dabei bezeichnet β_{jk} das Gewicht der direkten Verbindung zwischen verdecktem Neuron y_j und dem Ausgabeneuron z_k. Auch hier wird, um zur Entscheidungsfindung zu

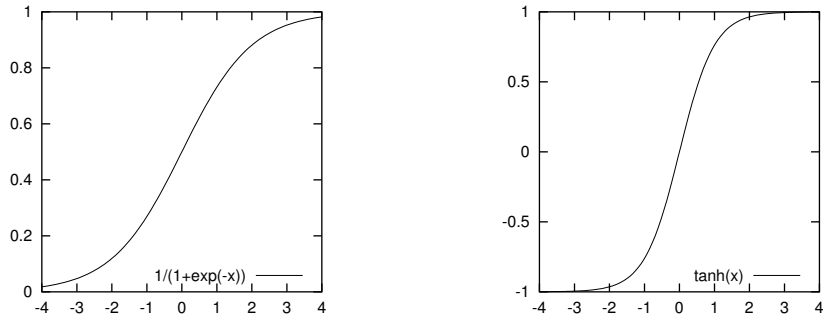


Abbildung 3.1: Von links nach rechts: Logistische und tangens–hyperbolicus Aktivierungsfunktion

gelangen, eine Aktivierungsfunktion benötigt. Wir verwenden ein zweites Mal die logistische Funktion und erhalten folgenden Ausdruck:

$$f(z_k) \quad = \quad f(\sum_{j=1}^{m} \beta_{jk} f(y_j)) = f(\sum_{j=1}^{m} \beta_{jk} f(\sum_{i=1}^{n} x_i \alpha_{ij})) \tag{3.5}$$

Diese Netze werden aufgrund der Richtung der Informationsverarbeitung auch als Feed–Forward–Netze bezeichnet. Eine weitere häufig verwendete Aktivierungsfunktion ist die tangens–hyperbolicus Funktion, die ähnliche Eigenschaften wie die logistische Aktivierungsfunktion besitzt. Abbildung 3.1 zeigt die tangens–hyperbolicus Funktion.

$$f(x) \quad = \quad tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{3.6}$$

3.1.3 Lernalgorithmen – Der Backpropagation–Algorithmus

Dieser Abschnitt soll den Lernalgorithmen gewidmet sein, speziell dem (Error)-Backpropagation-Algorithmus (Fehlerrückführungsalgorithmus). Ziel der Lernphase ist es, die Gewichte α_{ij} und β_{jk} so zu ermitteln, dass die Anzahl der Fehlklassifizierungen minimiert wird. Dabei werden dem Netz Trainingsbeispiele mit bekannter Klassenzugehörigkeit zur Verfügung gestellt. In der Trainingsphase soll durch optimale Wahl der Gewichte die Differenz zwischen Netzwerkklassifikation und dem zu lernenden Zielwert minimiert werden, daher der Name Fehlerrückführungsalgorithmus. Zur Erläuterung des Algorithmus sei beispielhaft eine neuronale Netzwerkarchitektur eines Feed–Forward–Netzes, bestehend aus Eingabeneuronen, inneren (verdeckten) Neuronen und einem Ausgabeneuron gegeben.
Das Flussdiagramm in Abbildung 3.2 stellt die Informationsverarbeitung eines neuronalen Netzes dar [Zim94].

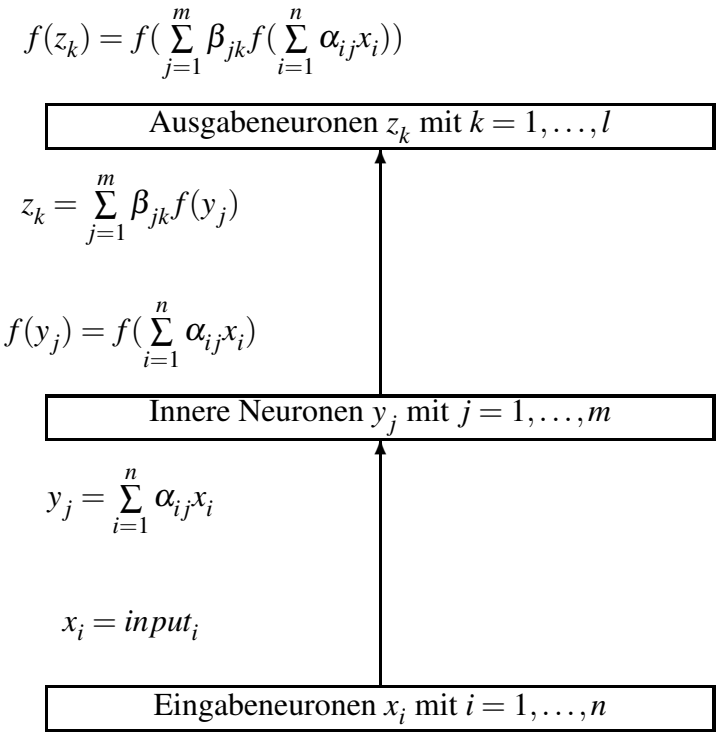


Abbildung 3.2: Informationsfluss in einem Feed–Forward–Netz

Dabei bezeichnen x_i, y_j und z_k den Nettoinput in die Neuronen der jeweiligen Neuronenschicht, $f(y_j)$ und $f(z_k)$ kennzeichnen die Ausgabe der Neuronen. Als Aktivierungsfunktion f wird beispielhaft die logistische Funktion gewählt. Schichtenweise werden in diesem Netz Informationen weiter transportiert. Dazu wird mittels der Aktivierungsfunktion die Aktivität eines Neurons bestimmt und an die nächste Schicht weitergegeben. Für nachfolgende Neuronen stellen die Aktivitäten voriger Neuronen somit Eingangssignale dar. In einem ersten Schritt wird das Ausgangssignal anhand vorher festgelegter Gewichte[1] ermittelt und für jedes Trainingsbeispiel t mit der Zielgröße verglichen. In weiteren Schritten sollen die Gewichte α_{ij} und β_{jk} so bestimmt werden, dass die mittlere quadratische Abweichung über alle Trainingsmuster zwischen Endausgabe $f(z_k)$ und Zielwert t_k minimiert wird. Lernen heißt also in diesem Sinne Fehlerminimierung. Der Einfachheit halber bezeichnen wir $f(z_k)$ als o_k, $f(y_j)$ als o_j und x_i als o_i. Dabei signalisiert o, dass es sich um eine Ausgabe (=Output) handelt. Daraus ergibt sich folgende Zielfunktion des Netzwerkfehlers $E(\alpha, \beta)$ über alle

[1] Der Initialwert jedes Gewichtsfaktors ist i. d. R. eine kleine Zufallszahl, z. B. aus dem Bereich -0.50 bis +0.50.

Trainingsbeispiele T:

$$E(\alpha,\beta) = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{l}\frac{1}{2}(o_k - t_k)^2 \longrightarrow min \tag{3.7}$$

bzw. für ein Trainingsbeispiel

$$E(\alpha,\beta) = \sum_{k=1}^{l}\frac{1}{2}(o_k - t_k)^2 \longrightarrow min. \tag{3.8}$$

Der Proportionalitätsfaktor von $\frac{1}{2}$ in Formel 3.7 dient der einfachen Herleitung der Lernregel. Bei Verwendung der logistischen Funktion bzw. der tanges–hyperbolicus Funktion gestaltet sich die Ermittlung der Ableitungen besonders einfach, da

$$f'(x) = f(x)(1 - f(x)) \tag{3.9}$$

bzw. für die tanges–hyperbolicus Funktion:

$$g'(x) = 1 - (g(x))^2 \tag{3.10}$$

gilt. Wir bestimmen zur Optimierung von 3.8 die folgenden Ableitungen unter Benutzung der logistischen Funktion als Aktivierungsfunktion. Dabei bezeichnen $\Delta\beta_{jk}$ und $\Delta\alpha_{ij}$ die Änderungsgrößen für die Gewichte β_{jk} und α_{ij} [Füs95].

1. Fehlerveränderung bei Änderung der Aktivität eines Ausgabeneurons o_k:

$$e_k = (o_k - t_k) \tag{3.11}$$

2. Fehlerveränderung bei Änderung des Nettoinputs z_k an einem Ausgabeneuron k:

$$\delta_k = e_k o_k (1 - o_k) \tag{3.12}$$

3. Fehlerveränderung bei Änderung des Gewichts β_{jk} zu einem Ausgabeneuron:

$$\Delta\beta_{jk} = \delta_k o_j \tag{3.13}$$

4. Fehlerveränderung bei Änderung der Ausgabe o_j eines Neurons der vorigen Ebene:

$$e_j = \sum_{k=1}^{l}\delta_k\beta_{jk} \tag{3.14}$$

5. Fehlerveränderung bei Änderung des Nettoinputs y_j eines Neurons der vorigen Ebene:

$$\delta_j = \sum_{k=1}^{l} \delta_k \beta_{jk} o_j (1 - o_j) \tag{3.15}$$

6. Fehlerveränderung bei Änderung des Gewichts α_{ij} zu einem inneren Neuron:

$$\Delta\alpha_{ij} = \delta_j o_i \tag{3.16}$$

7. Fehlerveränderung bei Änderung der Ausgabe eines Eingabeneurons x_i:

$$e_i = \delta_j \alpha_{ij} o_i \tag{3.17}$$

Abbildung 3.3 zeigt exemplarisch den vollständigen Ablauf von Informationsfluss und Fehlerausbreitung in einem Feed–Forward–Netz. Beginnend von der Ausgabeebene wird rückwärts der Einfluss der Ebenen auf den Fehler $E(\alpha, \beta)$ ermittelt. Bei Schritt drei und sechs wurde die Veränderung des Fehlers bei Veränderung der Gewichte β und α ermittelt. Der Fehler wird durch das allmähliche Anpassen der Gewichte reduziert. Dieser Prozess wird als Lern- bzw. Trainingsprozess bezeichnet, in dem die einzelnen Schritte zyklisch wiederholt werden.

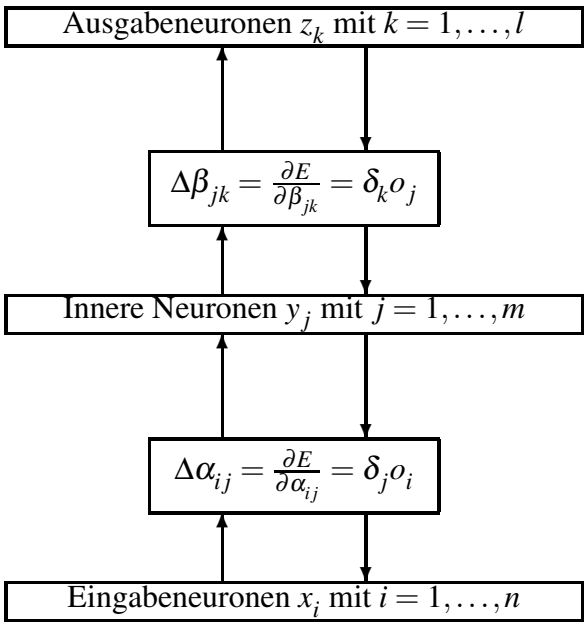


Abbildung 3.3: Informationsfluss und Fehlerausbreitung in einem Feed–Forward–Netz

3.1.4 Optimierungsverfahren

Der Backpropagation–Algorithmus liefert ein Verfahren zur Berechnung der ersten Ableitung der Zielfunktion. Damit ist noch nicht festgelegt, wie die Gewichte verändert werden sollen um eine optimale Minimierung der Zielfunktion zu erreichen. Alle Optimierungsverfahren wählen einen geeigneten zufälligen Startwert der Gewichte. Danach wird die Suchrichtung (Richtungsvektor) d im hochdimensionalen Raum der Gewichte bestimmt, entlang der man eine Verbesserung der Zielfunktion erwartet. In einem zweiten Schritt wird die Schrittweite (Lernrate) η in die Abstiegsrichtung festgelegt. Daraus erhält man einen neuen Parameterwert w_{n+1}. Diese Vorgehensweise wird solange iteriert, bis bestimmte Abbruchkriterien erfüllt sind [And97]. Im Folgenden werden wir alle Gewichte β und α zu einem Parameterwert w zusammenfassen. Die Gewichtsveränderung kann also folgendermaßen beschrieben werden:

$$w_{n+1} = w_n + \eta d \tag{3.18}$$

Von einem Gradientenverfahren spricht man, wenn als Suchrichtung d für die Minimierung der negative Gradient der Fehlerfunktion verwendet wird. Dabei bestimmt der Gradient an einer beliebigen Stelle der Fehlerfunktion den steilsten Anstieg. Es existieren weitere Verfahren die nicht den Gradienten verwenden, z. B. Simulated–Annealing oder genetische Algorithmen, auf die wir aber nicht näher eingehen wollen. Eine ausführliche Darstellung dieser Verfahren findet sich z. B. unter [MD89]. Wir bezeichnen mit g den Gradienten und mit $E(w)$ die Abhängigkeit der Fehlerfunktion von den Gewichten w. Die Suchrichtung d kann folgendermaßen beschrieben werden:

$$d = -g = -\frac{\delta E(w)}{\delta w} \tag{3.19}$$

Das Gradientenverfahren erzeugt eine Folge $E(w_0), E(w_1), \ldots, E(w_n)$ mit der Eigenschaft $E(w_{n+1}) \leq E(w_n)$. Entwickelt man die Taylorreihe der Zielfunktion bis zur zweiten Ordnung erkennt man sofort, dass durch eine Verschiebung in Richtung des negativen Gradienten eine Verkleinerung des Fehlers bewirkt wird [Zim94].

$$\begin{aligned} E(w_{n+1}) &= E(w+\eta d) \\ &= E(w-\eta g) \\ &\cong E(w) + \frac{\delta E(w)}{\delta w}(-\eta g) + \frac{1}{2}(-\eta g)\frac{\delta^2 E(w)}{\delta w^2}(-\eta g) \\ &\cong E(w) - \eta\langle g,g\rangle + \frac{\eta^2}{2} g \frac{\delta^2 E(w)}{\delta w^2} g \end{aligned} \tag{3.20}$$

Wird die Lernrate η genügend klein gewählt, so verliert der letzte Term durch die Multiplikation mit η^2 schnell an Bedeutung. Da das Skalarprodukt $\langle g,g\rangle$ nicht negativ ist, wird die Ziel- bzw. Fehlerfunktion nicht vergrößert.
Das Gradientenverfahren wird meist in erweiterter Form eines konjugierten Gradientenverfahrens angewendet. Dabei wird zusätzlich eine Schrittweitensteuerung und ein

Gedächtnisterm κ eingeführt. Der Gedächtnisterm erlaubt es, vergangene Suchrichtungen d in die Berechnung des neuen Parameters w_{n+1} miteinzubeziehen. Die Suchrichtung ergibt sich dann aus einer Überlagerung der Richtung des steilsten Abstieges und einer Dämpfung der alten Suchrichtung $d_{n+1} = -g + \kappa d_n$ mit $(0 < \kappa < 1)$. Weitere Informationen zum konjugierten Gradientenabstiegsverfahren finden sich u. a. in [And97].

3.1.5 Pruning–Algorithmen zur Ermittlung einer Basisnetzwerkarchitektur

Die Netzwerkarchitektur ist dafür verantwortlich wie gut die Zielfunktion approximiert werden kann. Die Zahl der inneren Neuronen und verwendeten Gewichte ist von entscheidender Bedeutung. Werden zu viele innere Neuronen verwendet, wird das Rauschen der Daten abgebildet. Eine häufig zitierte Daumenregel schlägt vor, als Anzahl der inneren Neuronen den Mittelwert der Anzahl der Ein- und Ausgabeneuronen zu wählen. Darüber hinaus könnten Hypothesentests für die einzelnen Gewichte durchzuführt werden. Gewichte, die nicht signifikant zur Erklärung der unabhängigen Variablen beitragen, könnten vernachlässigt und durchtrennt werden.

Eine weitere Möglichkeit zur Lösung dieses Problems bieten sogenannte Pruning–Algorithmen. Ziel der Pruning–Algorithmen ist es, eine geeignete Netzwerkarchitektur aufzustellen. Dabei ist das Vorgehen ähnlich wie bei Hypothesentests, allerdings wird anstelle der Prüfstatistik die sogenannte Saliency ermittelt, die ein Maß für die Bedeutung des Gewichts darstellt. Die Saliency eines Gewicht stellt die Zunahme des Netzwerkfehlers bei Entfernung des Gewichts aus dem Netzwerk dar. Die Herleitung des Verfahrens und die Ermittlung der Saliency beruht auf einer Approximation der Fehlerfunktion durch ein Taylorpolynom zweiter Ordnung (siehe 3.20) in der Umgebung der geschätzen Gewichte und ist in [LDS90, Zim94, And97] ausführlich dargestellt. Daneben werden in [LDS90, HS93] Verfahren zur Ausdünnung der neuronalen Netze vorgestellt.

3.1.6 Verwendung von neuronalen Netzen

Neuronale Netze besitzen eine hohe Lernfähigkeit und sind in der Lage auch nichtlineare Zusammenhänge in den Daten zu erkennen. Sie können „wichtige" Informationen aus den Datensätzen herausfiltern, indem sie Eingabeneuronen, die keinen Beitrag zur Klassifikation leisten, vom Netz abtrennen. Zudem unterscheidet sich die Art der Modellierung wesentlich von der Modellierung logischer Modelle. Aus diesen Gründen wollen wir die neuronalen Netze auf unsere Problemstellung anwenden und die Ergebnisse mit den Ergebnissen der logischen Modelle vergleichen.

3.2 Logisches Modell von Truemper

3.2.1 Grundidee

Das logische Klassifikationsmodell von Truemper [Tru04] zählt ebenfalls zu den überwachten Lern- bzw. Klassifikationsverfahren. In den folgenden Ausführungen gehen wir wieder von einer zweiklassigen Problemstellung aus. Die Grundidee des Modells besteht darin, dass aus den Daten logische Formeln extrahiert werden, welche die Datensätze separieren. Dies wird auch als logisches Lernen bezeichnet. Man sucht eine logische Formel D, die für die Klasse A stets *wahr* und für die Klasse B stets *falsch* liefert. Die logische Formel D separiert dann die Klasse A von B. Der Vorteil dieser Modellierung liegt im logischen Ansatz und der daraus resultierenden besseren Interpretier- und Nachvollziehbarkeit der Ergebnisse.

3.2.2 Voraussetzungen

Da es sich um ein überwachtes Klassifikationsverfahren handelt, ist die Kenntnis der Klassenzugehörigkeit der Datensätze Voraussetzung. Weiterhin ist es notwendig, dass die Daten in binärer Form vorliegen, d. h. jedes Merkmal muss in der Form: „Merkmal ist vorhanden: ja oder nein" vorliegen. Dies stellt jedoch bei allen logischen Klassifikatoren eine große Hürde dar, da die Daten naturgemäß nicht diese Gestalt aufweisen. Vielmehr liegen die Daten häufig in kategorischer (z. B. Beruf, Tarif, ...) oder rationaler Form (z. B. Darlehenshöhe, WoP–Höhe, ...) vor, die in binäre Form transformiert werden müssen. Zur Datentransformation existieren verschiedene Methoden. Quinlan verwendet z. B. bei der Klassifikation mit Hilfe von Entscheidungsbäumen einen entropiebasierten Ansatz aus der Informationstheorie [Qui93]. Dieser Ansatz wird im folgenden Kapitel ausführlich dargestellt.
Im Modell von Truemper wird ein anderer Ansatz gewählt. Es wird versucht, die rationalen Einträge in Intervalle einzuteilen, die dann entsprechend kodiert werden: „Eintrag y_l liegt im Intervall $[x_i, x_j]$ ja oder nein". Diese Transformation unterscheidet sich für kategorische und rationale Daten.

3.2.3 Transformation der Daten

3.2.3.1 Transformation von kategorischen Daten

Für kategorische Daten gestaltet sich die Transformation recht einfach, da die Anzahl der möglichen Einträge in den meisten Fällen überschaubar ist. In diesem Fall werden gerade so viele Variablen wie mögliche Merkmalsausprägungen bereitgestellt. Beispielsweise umfasst die Kategorie Tarif vier mögliche Merkmalsausprägungen, daher werden die Variablen x_1, x_2, x_3 und x_4 erzeugt. Stammt ein Konto aus Tarif 1, so wird dies folgendermaßen kodiert $x_1 = 1$, $x_2 = x_3 = x_4 = 0$. Für die drei weiteren Tarife wird ebenso verfahren.

3.2.3.2 Transformation von rationalen Daten

Die Transformation von rationalen Daten im Modell von Truemper erfolgt in mehreren Schritten. Die Grundidee der Methode besteht darin, nach abrupt wechselnden Klassenzugehörigkeiten zu suchen und diese als natürliche Intervallgrenzen zu wählen. Dazu werden die rationalen Einträge aufsteigend sortiert und mit ihrer Klassenzugehörigkeit markiert. In einem ersten Schritt werden dann die Werte (Schnittpunkte) gesucht, an denen sich die Klassenzugehörigkeit ändert. Dabei werden bis zu drei Schnittpunkte als Initialisierungspunkte gewählt, die auf natürliche Weise vier Intervalle definieren und eine binäre Kodierung erlauben. Die Initialisierungspunkte werden folgendermaßen gewählt:
Durchsuche alle rationalen Einträge v_i nach einem Wechsel der Klassenzugehörigkeit. Wird ein Wechsel gefunden, so wird die Differenz δ_i ermittelt.

$$\delta_i = |v_i - v_{i-1}|$$

Als Initialisierungsschnittpunkte z_i werden die Werte mit dem höchsten δ_i gewählt. Die z_i werden folgendermaßen ermittelt:

$$z_i = \frac{v_{i-1} + v_i}{2}$$

Mit Hilfe der z_i wird eine erste Transformation in logische Daten durchgeführt und überprüft, ob die Darstellung ausreichend präzise ist um daraus logische Formeln zu extrahieren. Dazu werden Teile eines Extrahierungsprozesses angestoßen, der in den folgenden Kapiteln dargestellt wird. Es wird überprüft, ob die „Clash-Bedingung" (diese Bedingung sorgt dafür, dass für alle Datensätze der Klasse *B falsch* geliefert wird) eingehalten werden kann, und damit die Existenz einer trennenden Formel gewährleistet ist. Die „Clash–Bedingung" wird ebenfalls im folgenden Abschnitt näher erläutert. Falls die Bedingung mit den gewählten z_i nicht eingehalten werden kann, wird ein rekursiver Prozess gestartet, der die Ausgangsintervalle weiter zerlegt. Die zusätzlichen Schnittpunkte werden wiederum mit Hilfe des Wechsels der Klassenzugehörigkeit bestimmt. Anschließend wird ein weiteres Mal überprüft, ob die Bedingung eingehalten werden kann. Ist dies der Fall, so bricht der Transformationsprozess ab, andernfalls werden die Intervalle rekursiv weiter verkleinert und zusätzliche Intervallschnittpunkte eingeführt.
Die Methode wird als „Cutpoint–Methode" bezeichnet und wurde von Bartnikowski, Granberry, Mugan und Truemper auch experimentell untersucht und lieferte robuste logische Datensätze, aus denen logische Formeln extrahiert werden konnten. Weitere Ausführungen zum „Cutpoint–Verfahren" und Anwendungsbeispiele finden sich in [BGMT04]. Das Verfahren ist auch im frei zugänglichen LEIBNIZ–System[2] zur Klassifikation implementiert.

[2]Das Softwaretool Leibniz kann kostenlos unter `http://utdallas.edu/~klaus/Leibnizprogram/leibnizmain.html` heruntergeladen werden.

3.2.3.3 Fehlende Werte

Im logischen Modell von Truemper werden zwei Klassen von fehlenden Werten unterschieden. Es sind die Einträge *fehlend* und *nicht ermittelbar* möglich. *Nicht ermittelbar* bedeutet, dass die Werte z. B. bei Patienten aufgrund bestimmter gesundheitlicher Konstellationen nicht gemessen werden können. Dies verdeutlicht, dass der nicht vorhandene Eintrag vom Datensatz abhängig ist und damit in gewisser Weise Informationen beinhaltet. Hingegen ist der Eintrag *fehlend* vom Datensatz unabhängig und beinhaltet keine Information. Es ist z. B. möglich, dass eine Messung am Patienten vergessen worden ist. Anwendungsbeispiele zeigten [MT04], dass diese Unterscheidung sinnvoll ist. Damit sind im Modell von Truemper folgende logische Einträge möglich: *wahr*, *falsch*, *nicht ermittelbar* und *fehlend*.
Sei r ein Datensatz mit den Variablen $x_1, x_2, \ldots, x_n$ und D_1 eine DNF–Klausel der Gestalt $D_1 = (x_1 \wedge \ldots \wedge x_n)$. Die Variablen können die Werte *wahr*, *falsch*, *nicht ermittelbar* oder *fehlend* besitzen. Die DNF–Klausel D_1 liefert *wahr*, wenn alle entsprechenden *wahr/falsch* Einträge im Datensatz r dazu führen, dass D_1 *wahr* liefert. Die Klausel D_1 liefert *falsch*, wenn eine entsprechende Variable im Datensatz *r falsch* liefert oder ein Literal in D_1 existiert, dessen Ausprägung in *r nicht ermittelbar* ist. Liefert D_1 gemäß diesen Definitionen nicht *wahr* oder *falsch*, so wird *nicht entscheidbar* als Wahrheitswert geliefert. Dies ist allerdings nur möglich, wenn mindestens ein Eintrag in r mit der Ausprägung *fehlend* vorhanden ist.
Diese Vorgehensweise kann auch mit intuitiven Argumenten begründet werden.

Beispiel 3.1. *Die Datensätze p_1, p_2 zweier Patienten seien mit:*

$$\begin{aligned} p_1 &: x_1 = \textit{nicht ermittelbar},\ x_2 = \textit{wahr},\ x_3 = \textit{falsch}, \ldots, x_n = \textit{wahr} \\ p_2 &: x_1 = \textit{fehlend},\ x_2 = \textit{wahr},\ x_3 = \textit{wahr}, \ldots, x_n = \textit{wahr} \end{aligned}$$

gegeben. Dabei sind die Ausprägungen der Werte $x_2, \ldots, x_n$ bei beiden Patienten vollständig bekannt. Der Wert x_1 kann beim ersten Patienten p_1 aus gesundheitlichen Gründen nicht ermittelt werden, beim Patient p_2 wurde keine bzw. eine fehlerhafte Messung durchgeführt. Gemäß obiger Definition kann D_1 nur wahr liefern, wenn der Wert für x_1 bekannt ist. Für Patient p_1 liefert D_1 aufgrund des nicht ermittelbaren Wertes stets falsch. Für Patient p_2 liefert D_1 hingegen den Wahrheitswert nicht entscheidbar. Dies ist auch intuitiv richtig, da die Messung für Patient p_2 wiederholt werden kann, und somit ein Wahrheitswert für D_1 ermittelt werden könnte. Eine wiederholte Messung ist für Patient p_1 nicht möglich, daher kann D_1 für p_1 nie den Wahrheitswert wahr liefern.

3.2.4 Formelermittlung

Wie erhält man bei gegebenen Datenmengen der Klassen A und B eine DNF–Klausel D_i, die für eine maximale Teilmenge A_i aus A *wahr* liefert und für alle Datensätze aus

B falsch erzeugt? Die allgemeine Form der DNF–Klausel D_i hat folgende Gestalt:

$$D_i = \bigwedge_{j=1}^{n} (x_j \text{ oder } \neg x_j \text{ oder kein Literal } x_j) \tag{3.21}$$

Zur Generierung der Formel werden für $j = 1,\ldots,n$ zwei logische Variablen $x_j(neg)$ und $x_j(pos)$ eingeführt, die zu folgender Auswahl der Literale für D_i führen:

$$\begin{aligned}
x_j(pos) &= \textit{wahr} \text{ und } x_j(neg) = \textit{falsch}\text{: Wähle Literal } x_j \\
x_j(pos) &= \textit{falsch} \text{ und } x_j(neg) = \textit{wahr}\text{: Wähle Literal } \neg x_j \\
x_j(pos) &= x_j(neg) = \textit{falsch}\text{: Wähle kein Literal} \\
x_j(pos) &= x_j(neg) = \textit{wahr}\text{: Nicht erlaubt}
\end{aligned}$$

Der vierte Fall wird mit folgender Bedingung ausgeschlossen:

$$\neg x_j(pos) \vee \neg x_j(neg), \quad j = 1,\ldots,n$$

Diese Bedingung besagt, dass $x_j(pos)$ und $x_j(neg)$ nicht gleichzeitig *wahr* liefern können. Sei r ein beliebiger Datensatz aus A oder B, so definieren wir $J_+^r = \{j \mid x_j = \textit{wahr} \text{ in Datensatz r}\}$, $J_-^r = \{j \mid x_j = \textit{falsch} \text{ in Datensatz r}\}$, $J_u^r = \{j \mid x_j = \textit{nicht ermittelbar} \text{ in Datensatz r}\}$ und $J_a^r = \{j \mid x_j = \textit{fehlend} \text{ in Datensatz r}\}$. Der Datensatz r kann daher folgendermaßen beschrieben werden:

$$\begin{aligned}
x_j &= \textit{wahr}, \ \forall j \in J_+^r \\
x_j &= \textit{falsch}, \ \forall j \in J_-^r \\
x_j &= \textit{nicht ermittelbar}, \ \forall j \in J_u^r \\
x_j &= \textit{fehlend}, \ \forall j \in J_a^r
\end{aligned}$$

Damit können Bedingungen formuliert werden, die von allen Datensätzen aus A und B erfüllt werden. Wir beginnen mit den Datensätzen der Menge B. Die Klausel D_i muss für alle Datensätze s aus *B falsch* liefern. Dies ist genau dann der Fall, wenn D_i das Literal x_j enthält, Datensatz s aber die Ausprägung $x_j = falsch$ besitzt bzw. D_i das Literal $\neg x_j$ enthält, Datensatz s aber die Ausprägung $x_j = \textit{wahr}$ aufweist. Weiterhin würde *falsch* erzeugt werden, wenn D_i die Literale x_j oder $\neg x_j$ enthält, Datensatz s aber das Merkmal $x_j = \textit{nicht ermittelbar}$ besitzt.
In logischer Formulierung sieht diese Bedingung, die im Modell von Truemper als „Clash–Bedingung" bezeichnet wird, folgendermaßen aus:

$$\bigvee_{j \in (J_+^s \cup J_u^s)} x_j(neg) \vee \bigvee_{j \in (J_-^s \cup J_u^s)} x_j(pos), \quad \forall s \in B \tag{3.22}$$

Die Klausel D_i soll für eine maximale Teilmenge $A_i \subseteq A$ *wahr* liefern, d. h. alle Literale müssen *wahr* liefern. Dies ist genau der Fall wenn, für $j \in (J_+^r \cup J_u^r \cup J_a^r)$, $x_j(neg) =$

$falsch$, und für $j \in (J^r_- \cup J^r_u \cup J^r_a)$, $x_j(pos) = falsch$ ist. Diese Bedingung wird als „Agree–Bedingung" bezeichnet und besitzt in logischer Formulierung folgende Gestalt:

$$\bigwedge_{j \in (J^r_+ \cup J^r_u \cup J^r_a)} \neg x_j(neg) \wedge \bigwedge_{j \in (J^r_- \cup J^r_u \cup J^r_a)} \neg x_j(pos), \quad \forall r \in A \tag{3.23}$$

Wir modellieren Zugehörigkeit von r in A_i mit einer logischen Variablen *select(r)* die *wahr* ist, wenn $r \in A_i$ bzw. *falsch*, wenn $r \notin A_i$ ist.

$$select(r) \rightarrow [\bigwedge_{j \in (J^r_+ \cup J^r_u \cup J^r_a)} \neg x_j(neg) \wedge \bigwedge_{j \in (J^r_- \cup J^r_u \cup J^r_a)} \neg x_j(pos)], \quad \forall r \in A \tag{3.24}$$

Die Klausel D_i kann folgendermaßen ermittelt werden: Ordne jedem *select(r)*, $r \in A$, Kosten in Höhe von 1 zu falls *select(r) falsch* liefert und 0 falls *select(r) wahr* liefert. Nun kann die MINSAT–Instanz, die durch die „Clash– und Agree–Bedingungen" und die Kosten von *select(r)* definiert ist, gelöst werden. Die im LEIBNIZ–System implementierten Lösungsalgorithmen entstammen unter anderem aus [Rem01]. Dort findet man auch eine ausführliche Darstellung der Lösungsalgorithmen. Die erfüllende Belegung entspricht unserer gesuchten logischen Formel D_1 die *wahr* für A_1 und *falsch* für B liefert.
Die weiteren Formeln findet man mit Hilfe eines rekursiven Prozesses. Unter der Annahme dass A_1 nicht leer ist, ersetzt man A durch $A \backslash A_1$ und bezeichnet die neue Menge mit A_2, die daraus resultierende DNF–Klausel mit D_2. Falls A_2 leer ist sind wir fertig und $D = D_1$. Ansonsten wird der Prozess fortgesetzt und wir erhalten D_2. Anschließend wird A durch $A \backslash (A_1 \cup A_2)$ ersetzt. Die Schritte werden solange wiederholt bis $A = \emptyset$ und wir erhalten für D:

$$D = D_1 \vee D_2 \vee \ldots \vee D_l \tag{3.25}$$

D ist damit *wahr* für alle Datensätze aus A und *falsch* für alle Datensätze aus B. Eine ausführliche Darstellung der Problemformulierung findet sich in [Tru04].

3.2.5 Existenz einer trennenden Formel

Eine trennende DNF–Formel existiert stets, wenn die Datensätze aus A und B nicht ineinander verschachtelt bzw. identisch sind. Ein Datensatz $r \in A$ ist schwach in einen Datensatz $s \in B$ verschachtelt, wenn für jede *wahr* oder *falsch* Ausprägung des Datensatzes r, der Datensatz s dieselbe *wahr* oder *falsch* Ausprägung bzw. *fehlend* besitzt.

Beispiel 3.2. *[Tru04] Die Menge A enthält den Datensatz r:*

$$r \;:\; x_1 = wahr,\ x_2 = falsch,\ x_3 = nicht\ ermittelbar$$

Die Menge B enthält den Datensatz s:

$$s \quad : \quad x_1 = wahr,\ x_2 = fehlend,\ x_3 = wahr$$

Für s ∈ B lautet die „Clash–Bedingung":

$$x_1(neg) \vee x_3(neg)$$

Allerdings wird damit schon deutlich, dass dieses System nicht erfüllbar ist, da die „Agree–Bedingung" für r ∈ A gerade fordert:

$$\neg x_1(neg) \wedge \neg x_2(pos) \wedge \neg x_3(pos) \wedge \neg x_3(neg)$$

Vor der Ermittlung einer trennenden DNF–Formel müssen die Daten hinsichtlich ihrer Eindeutigkeit untersucht und gegebenenfalls verschachtelte Datensätze entfernt werden. Das folgende Theorem zeigt, dass unter Ausschluss der schwachen Verschachtelung stets eine trennende Formel D existiert.

Theorem 3.1. *[Tru04] Seien A und B zwei Mengen von Datensätzen. Es existiert eine trennende DNF–Formel D, die A von B separiert, genau dann wenn kein Datensatz r ∈ A schwach in einen Datensatz s ∈ B verschachtelt ist.*

Beweis. Angenommen es existiert eine DNF–Formel D die A von B separiert. Wähle einen beliebigen Datensatz $r \in A$. Für diesen Datensatz liefert D den Wert *wahr*. Daher muss D eine Klausel D_k besitzen, dass für jedes Literal x_j bzw. $\neg x_j \in D_k$, der Datensatz $r \in A$ $x_j = wahr$ bzw. $x_j = falsch$ besitzt. Für jeden beliebigen Datensatz $s \in B$ muss D_k den Wert *falsch* liefern. Daher müssen Literale in D_k, z. B. mit Index j existieren, für die der Datensatz $s \in B$ $x_j = nicht\ ermittelbar$ aufweist. Oder aber der Literal x_j ist in D_k enthalten und s enthält $x_j = falsch$ oder Literal $\neg x_j$ ist in D_k vorhanden und s besitzt $x_j = wahr$. Daraus folgt, dass bei Existenz einer trennenden DNF–Formel der Datensatz r nicht schwach in s verschachtelt sein kann.

Im zweiten Fall gehen wir von der Annahme aus, dass kein $r \in A$ schwach in ein beliebiges $s \in B$ verschachtelt ist. Wir definieren D als DNF–Formel, in der jede DNF–Klausel D_r durch ein $r \in A$ folgendermaßen definiert ist. Für jedes $x_j = wahr$ bzw. $x_j = falsch$ in r, enthält D_r den Literal x_j bzw. $\neg x_j$. Mit dieser Konstruktion liefert D_r den Wert *wahr* für r. Wähle nun ein beliebiges $s \in B$. Da r nicht schwach nicht s verschachtelt ist, besitzt r einige Literale x_j mit *wahr* oder *falsch* Einträgen, in denen s gegenteilige *wahr* oder *falsch* Werte aufweist bzw. den Eintrag *nicht ermittelbar* besitzt. Egal um welchen Fall es sich handelt, die Klausel D_r liefert *falsch* für s. Daraus schließen wir, dass die DNF–Formel D, die aus den Klauseln D_r, $r \in A$ besteht, für alle Datensätze aus A den Wert *wahr* liefert und für alle Datensätze aus B den Wert *falsch* zurückgibt. □

3.2.6 Wert und Bewertung einer Formel

Wir wollen uns in den folgenden Ausführungen mit der Bewertung der DNF–Formel bzw. der Bewertung der DNF–Klauseln beschäftigen. Die in Unterabschnitt 3.2.4 erzeugte DNF–Formel ist von der Gestalt $D = D_1 \vee D_2 \vee \ldots \vee D_n$, mit $D_n = (x_i \wedge x_j \wedge \ldots \wedge x_m)$. Eine Klausel dieser DNF liefert den Wahrheitswert *wahr*, wenn alle Literale der DNF–Klausel den Wahrheitswert *wahr* liefern. Die Klausel liefert den Wahrheitswert *falsch*, sobald ein Literal der Klausel den Wahrheitswert *falsch* erzeugt.
Damit die DNF–Formel den Wahrheitswert *wahr* liefert, ist es ausreichend, wenn eine der DNF–Klauseln den Wahrheitswert *wahr* erzeugt. Die DNF–Formel liefert hingegen den Wahrheitswert *falsch*, wenn alle DNF–Klauseln der Formel den Wert *falsch* liefern.
Die Bewertung der DNF–Formel erfolgt in Abhängigkeit der Klassenzugehörigkeit des Datensatzes, wie in Tabelle 3.1 dargestellt. Wenn der Wert einer Formel den Wahr-

Wahrheitswert der Formel	Formel *wahr* → Stimme für	Bewertung
wahr	A	1
falsch	B	1
wahr	B	-1
falsch	A	-1

Tabelle 3.1: Formelbewertung im Modell von Truemper

heitswert *wahr* liefert und die Formel *wahr* = Stimme für A impliziert, so bekommt der Datensatz den Wert +1 zugewiesen. Die Formel deutet also darauf hin, dass der Datensatz aus Klasse *A* stammt. Liefert die Formel den Wahrheitswert *falsch*, so bekommt der Datensatz den Wert -1 zugewiesen. Dies bedeutet, dass der Datensatz vermutlich aus der Menge *B* stammt. Impliziert die Formel *wahr* = Stimme für B, so bekommt der Datensatz den Wert -1 zugewiesen, wenn der Wert der Formel den Wahrheitswert *wahr* liefert. Liefert die Formel den Wahrheitswert *falsch*, so erhält der Datensatz den Wert +1 zugewiesen.

Im LEIBNIZ–System von Truemper können insgesamt 40 Formeln generiert werden. Dabei garantiert jede Formel eine Teilmenge der Datensätze aus A und alle Datensätze aus B richtig zu klassifizieren. Für jeden Datensatz werden die Stimmen aller 40 Formeln addiert. Die Gesamtanzahl der Stimmen je Datensatz schwankt daher zwischen -40 bis 40. Dabei entspricht eine Gesamtanzahl von 40 Stimmen einem maximalen Grad der Zugehörigkeit zur Klasse *A*. Beträgt die Gesamtanzahl der Stimmen -40, so entspricht dies einem maximalen Grad der Zugehörigkeit zur Klasse *B*. Allerdings ist es im Modell von Truemper auch möglich, dass ein Datensatz 0 Stimmen erfüllt. Dieser Datensatz bekommt den Wert *nicht entscheidbar* zugewiesen. Nähere Informationen zur Formelgenerierung finden sich in [Tru04].

Mit Hilfe einer Testmenge wird die Klassifikationsgüte des Modells überprüft. Dabei wird die Stimmanzahl jedes Datensatzes aus der Testmenge gezählt und Sensitivität und Spezifität für verschiedene Schwellwerte aus dem Intervall $[-40, 40]$ ermittelt.

3.2.7 Verwendung des logischen Modells von Truemper

Das logische Klassifikationsmodell von Truemper versucht mit Hilfe einer DNF–Formel die Klasse *A* von der Klasse *B* zu separieren. In Kapitel 4 wird mit Hilfe der Implikationentheorie der formalen Begriffsanalyse ein Modell entwickelt, das versucht logische Regeln zu generieren. Die Zusammenfassung dieser Regeln liefert ebenfalls DNF–Formeln, welche die Klassen unserer Problemstellung trennen sollen. Aufgrund der Ähnlichkeiten wollen wir das Modell von Truemper für unsere Fragestellung verwenden. Zudem ist es uns möglich das generelle Abschneiden zweier logischer Klassifikationsmodelle direkt miteinander zu vergleichen.

3.3 Entscheidungsbäume

3.3.1 Grundidee

Die Entscheidungsbäume stellen ebenfalls ein überwachtes Lern- bzw. Klassifikationsverfahren dar. Dabei wird versucht, Klassifikationsregeln mit Hilfe einer Baumstruktur zu erzeugen. Die Bewertung und Auswahl der Klassifikationsmerkmale erfolgt mit Hilfe eines entropiebasierten Ansatzes.

3.3.2 Modellierung

Ein Entscheidungsbaum verlangt als Eingabe ein Objekt das durch mehrere Attribute beschrieben wird [RN04]. Ziel ist es, die Objekte anhand ihrer Attribute in Klassen einzuteilen. Dabei entspricht jeder interne Knoten im Entscheidungsbaum einem Test auf das Vorhandensein eines bestimmten Merkmals. Die Kanten, die von den Knoten abgehen, stellen dementsprechend die Testausgänge dar. Blattknoten beinhalten das Klassifikationsergebnis und bestimmen somit die Klassenzugehörigkeit des Pfades. Im Idealfall befinden sich nur Datensätze einer Klasse in einem Blatt. Soll beispielsweise das Kreditrisiko eines unbekannten Darlehensnehmers ermittelt werden, so durchwandert der zugrunde liegende Datensatz ausgehend von der Wurzel den gesamten Baum, bis er anhand verschiedener Tests an einem Blatt angelangt ist. Abbildung 3.4 zeigt einen solchen Entscheidungsbaum zur Abschätzung des Kreditrisikos. Allerdings soll der Baum dabei möglichst kompakt und übersichtlich bleiben. Er soll eine geringe Tiefe und daraus resultierend kurze Pfade aufweisen. Damit wird eine Abbildung des Rauschens, also ein Overfitting–Effekt vermieden. Um den Entscheidungsbaum kompakt zu halten, müssen „wichtige" Merkmale (Merkmale, die sehr viel Information

beinhalten) zur Klassifikation gewählt werden. Dadurch wird eine korrekte Klassifikation mit möglichst wenig Testinstanzen erreicht. Dies wirft allerdings die Frage auf, wie solche Merkmale ausgewählt bzw. deren Informationsgehalt ermittelt werden soll.

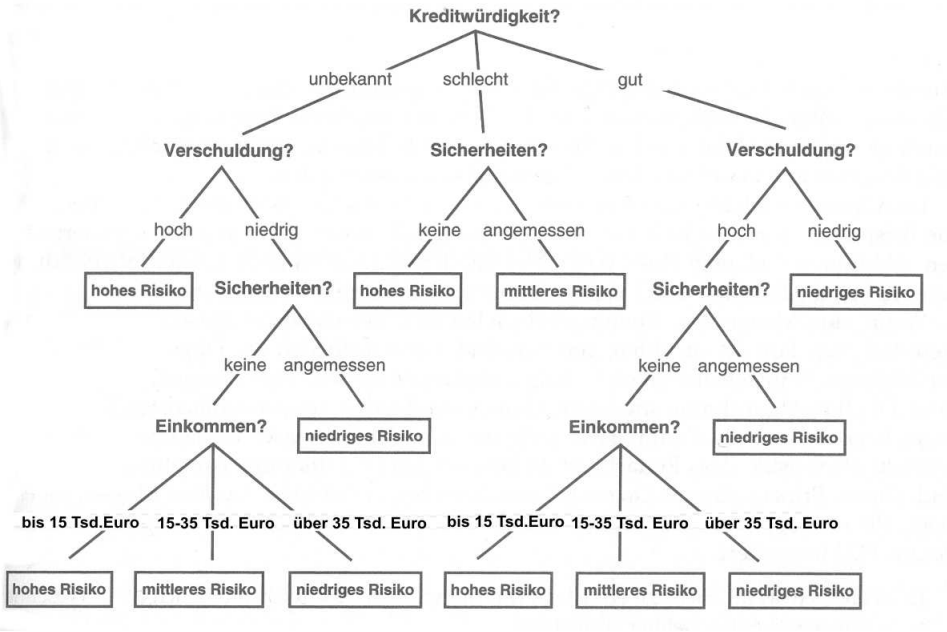


Abbildung 3.4: Entscheidungsbaum zur Klassifizierung des Kreditrisikos [Lug01].

3.3.3 Merkmalsauswahl

Wir suchen also nach einem Maß für die Attribute, das uns einen Entscheidungsbaum mit möglichst geringer Tiefe liefert. Ein perfektes Attribut würde die Grundgesamtheit so aufspalten, dass sich in jeder Menge nur noch Datensätze einer Klasse befinden. Ein schlechtes Attribut würde nach dem Test die Datensätze in den Mengen im gleichen Verhältnis wie in der Ausgangsmenge zurücklassen.
Ein geeignetes Maß hierfür ist die erwartete Informationsmenge die ein Attribut liefert. Der mathematische Begriff des Informationsgehaltes wurde in [SW49] definiert. Nach Shannon ist der Informationsgehalt einer Nachricht abhängig von der Wahrscheinlichkeit des Auftretens des Attributes. Shannon formalisierte dies, indem er den Informationsgehalt einer Nachricht als Funktion der Wahrscheinlichkeit p des Vorkommes jeder möglichen Nachricht als $-log_2 p$ definiert hat. Der Informationsgehalt einer Nachricht wird dabei in Bit gemessen. Mit einem Bit können zwei Zustände kodiert werden (z. B. Kopf oder Zahl). Daher ist ein Bit Information ausreichend um eine „Ja–Nein" Frage zu beantworten (z. B. Münzwurf).

Die Wahrscheinlichkeit des Auftretens eines Merkmales C_j wird folgendermaßen ermittelt [Qui93]:

$$\frac{freq(C_j,S)}{|S|}$$

Dabei bezeichnet S eine Menge von Datensätzen, $|S|$ die Anzahl dieser Datensätze und C_j eine beliebige Klasse (z. B. Kopf). Der Informationsgehalt wird dabei in Bits als –Logarithmus zur Basis 2 gemessen. Daher beträgt der Informationsgehalt der Klasse C_j

$$-log_2(\frac{freq(C_j,S)}{|S|}) \quad \text{Bits.}$$

Um den Informationsgehalt aller Datensätze aus S zu bestimmen, wird über alle Klassen C_j (z. B. Kopf und Zahl) summiert:

$$info(S) \quad = \quad -\sum_{j=1}^{k} \frac{freq(C_j,S)}{|S|} \cdot log_2(\frac{freq(C_j,S)}{|S|}) \quad \text{Bits}$$

Beispiel 3.3. *Es soll der Informationsgehalt eines Münzwürfes mit einer ungefälschten (fairen) Münze ermittelt werden.*

$$info(Münzwurf) \quad = \quad -\frac{1}{2}log_2(\frac{1}{2}) - \frac{1}{2}log_2(\frac{1}{2}) = \textit{ 1 Bit}$$

Die Größe $info(S)$ wird auch als Entropie der Menge S bezeichnet. Damit kann die gesamte Information ermittelt werden, die benötigt wird, um alle gegebenen Beispieldatensätze mit einem Entscheidungsbaum abzudecken.
Es stellt sich allerdings die Frage, welches Merkmal als Wurzelknoten verwendet werden soll. Welches Attribut bzw. welcher Test liefert zu Beginn die beste Aufspaltung? Dazu wird ein beliebiges Attribut als Testattribut gewählt und seine Testausgänge ermittelt. Es wird überprüft, wieviel Information nach Wahl des Wurzelknotens noch benötigt wird, um den Baum fertigzustellen. D. h. wieviel Information steckt nach Wahl des Wurzelknotens noch in den einzelnen Teilbäumen.
Es sei eine Trainingsmenge T gegeben. Wenn wir die Eigenschaft X mit n möglichen Ausgängen als Testknoten wählen, wird die Trainingsmenge dadurch in Teilmengen $T_1, T_2, \ldots, T_n$ zerlegt. Wird X als Wurzelknoten eingesetzt, so ist für die Fertigstellung des Baumes noch folgende Informationsmenge notwendig [Lug01]:

$$info_X(T) \quad = \quad \sum_{i=1}^{n} \frac{|T_i|}{|T|} \cdot info(T_i)$$

Die Größe $info_X(T)$ entspricht der noch benötigten Informationsmenge, in Form der gewichteten Summe aller entstandenen Teilbäume. Der Informationsgewinn durch die Wahl der Eigenschaft X als Testgröße berechnet sich dann folgendermaßen:

$$gain(X) \quad = \quad info(T) - info_X(T)$$

Bei einer algorithmischen Lösung des Problems (ID3–Algorithmus in [Qui93]) werden alle Attribute X auf ihren Informationsgewinn hin untersucht. Das Attribut mit dem höchsten Gewinn $gain(X)$ wird dann als Wurzelknoten gewählt. Der ID3–Algorithmus führt diese Analyse rekursiv mit jedem Teilbaum durch bis der Baum fertiggestellt ist.

3.3.4 Erweiterungen

Mit diesem Verfahren können einfache Entscheidungsbäume zur Klassifikation erstellt werden. Allerdings sind im ID3–Algorithmus folgende Probleme unberücksichtigt:

- Rationale Merkmale
  Wie sollen rationale Merkmalsausprägungen, wie z. B. WoP–Höhe, Einkommen, Darlehenshöhe, ... als Testgrößen behandelt werden?

- Fehlende Werte
  Jeder Test basiert auf einem beliebigen Merkmal. Wie soll klassifiziert werden, wenn dieses Merkmal bei einem Datensatz nicht vorhanden ist?

Diese Fragestellungen wurden in der Erweiterung des ID3–Algorithmus miteinbezogen. Der C4.5–Algorithmus[Qui93] ist in der Lage rationale Dateneinträge, fehlende Werte oder verrauschte Daten zu verarbeiten.
Bei rationalen Merkmalen wie Darlehenshöhe oder Spardauer werden die Trainingsdatensätze T bezüglich der rationalen Ausprägungen des Attributs X sortiert. Da es nur eine endliche Anzahl dieser Werte gibt, können diese folgendermaßen sortiert werden: $\{v_1, v_2, \ldots, v_m\}$ mit $v_1 \leq v_2 \leq \ldots \leq v_m$. Jeder Schnittpunkt trennt die Objekte in zwei Klassen. Zum einen Datensätze, die im Intervall $\{v_1, v_2 \ldots, v_i\}$ liegen und zum anderen Datensätze, die sich im Intervall $\{v_{i+1}, v_{i+2}, \ldots, v_m\}$ befinden. Es existieren genau $m-1$ mögliche Schnittpunkte, die im C4.5–Algorithmus alle hinsichtlich ihres Gewinns $gain(X)$ untersucht werden. Die Umsetzung ist aufgrund der vorab durchgeführten Sortierung relativ einfach. Die Schnittpunkte liegen genau im Intervallmittelpunkt [Qui93]:

$$\frac{v_i + v_{i+1}}{2}$$

Allerdings wird die benötigte Laufzeit um einen Entscheidungsbaum zu erstellen stark von der Sortierung dominiert. Die oben erwähnten Operationen können in linearer Zeit durchgeführt werden, die Sortierung von d rationalen Werten benötigt aber die Laufzeit $O(d \cdot \log(d))$ [Qui93].

Bei fehlenden Werten, die in der Realität häufig auftreten, ergeben sich zwei Probleme. Zum einen stellt sich die Frage, wie bei fehlenden Werten, die keine Aussage über einen Testausgang möglich machen, in der Trainingsmenge verfahren werden soll. Zum anderen können auch ungesehene Datensätze aus der Testmenge nicht klassifiziert werden. Die einfachste Möglichkeit bestünde darin, diese Datensätze für das Training nicht zu verwenden bzw. bei der Validierung das Ergebnis „nicht klassifizierbar" auszugeben. Allerdings wird dadurch die Anzahl der Trainings- und Testdaten reduziert. Da die Datenbasis in vielen Anwendungen häufig gering ist, erscheint das Vorgehen nicht praktikabel.
Daher wird ein probabilistischer Ansatz gewählt und die Ermittlung des $gain(X)$ dahingehend modifiziert, dass nur Datensätze verwendet werden, deren Merkmalsausprägung für das Attribut A bekannt ist. Der modifizierte $gain(X)$ ermittelt sich folgendermaßen:

$$gain(X) \quad = \quad F \cdot (info(T) - info_X(T))$$

Dabei bezeichnet F den Anteil der Datensätze, deren Ausprägung für das Attribut A bekannt ist. Zur Ermittlung des Informationsgehaltes des gesamten Baumes und eines Attributs A, werden ebenfalls nur Datensätze mit bekannten Werten verwendet. Den Datensätzen werden nun Gewichte zugewiesen. Kann ein Datensatz eindeutig einer Klasse zugewiesen werden, so erhält er das Gewicht 1. Eine solch starke Zuweisung ist bei fehlenden Einträgen nicht möglich. Wird ein Test X auf ein bestimmtes Attribut A mit $O_1, \ldots, O_n$ möglichen Ausgängen durchgeführt, so wird der Datensatz auf alle Testausgänge verteilt, und ihm wird die Wahrscheinlichkeit jedes Testausgangs als Gewicht zugewiesen.

Die Klassifikation ungesehener Fälle weist ein ähnliches Vorgehen auf. Der Datensatz wird den möglichen Ausgängen mit der Wahrscheinlichkeit des Auftretens des Ausgangs zugeordnet, er wird auf mehrere Pfade verteilt. An den Blättern angekommen wird die Klasse mit der höchsten Wahrscheinlichkeit gewählt.

3.3.5 Pruning von Entscheidungsbäumen

Bei der Modellierung mit neuronalen Netzen wurden Pruning–Methoden zur Gewichtsausdünnung verwendet. Auch die Entscheidungsbäume nutzen Pruning–Methoden zur „Entzweigung" bzw. Vereinfachung eines Baumes.
Im Idealfall wird die Generierung eines Entscheidungsbaumes so lange fortgesetzt, bis sich in jedem Blatt nur noch Datensätze einer Klasse befinden. Dadurch entstehen allerdings sehr komplexe Bäume, die häufig das Rauschen abbilden, also einen Overfitting–Effekt zur Folge haben.
Bei der Vereinfachung von Entscheidungsbäumen existieren grundsätzlich zwei Möglichkeiten. Zum einen besteht die Möglichkeit in einem Blatt die Entscheidung zu treffen, dass nicht mehr weiter aufgespalten werden soll. Zum anderen können bereits

erzeugte Teilbäume im Nachhinein wieder gelöscht werden.
Eine Konsequenz des Prunings ist allerdings, dass die Datensätze in einem Blatt nicht notwendigerweise derselben Klasse angehören müssen. Es wird also keine Klasse mehr mit einem Blatt identifiziert, sondern vielmehr eine Klassenverteilung, die für jedes Blatt eine Wahrscheinlichkeit spezifiziert, dass ein Datensatz dieses Blattes zu einer bestimmten Klasse gehört.

Entscheidungsbäume werden häufig dadurch vereinfacht, dass ein oder mehrere Teilbäume durch Blätter ersetzt werden [Qui93]. Dabei bezeichnet N die Anzahl der Datensätze in einem Blatt und E entspricht der Anzahl der fehlerhaften Datensätze in einem Blatt. Damit liefert $\frac{E}{N}$ die Fehlerrate eines Blattes.
Unter der Annahme, dass die Fehlerrate eines Baumes einschließlich seiner Teilbäume und aller Blätter vorhergesagt werden könnte, ergäbe sich eine relativ einleuchtende Pruning–Regel: Es wird von den Blättern aufwärts am Entscheidungsbaum damit begonnen, jeden Teilbaum zu untersuchen der kein Blatt ist. Falls die Ersetzung des Teilbaumes durch ein Blatt eine niedrigere Fehlerrate als die vorhergesagte Rate liefert, so wird der Teilbaum durch ein Blatt ersetzt. Anderenfalls bleibt der Teilbaum bestehen. Da die Fehlerrate in jedem Teilbaum reduziert wird, reduziert sich dadurch auch die Fehlerrate des kompletten Entscheidungsbaumes.
Allerdings stellt sich die Frage, wie die Fehlerraten eines Entscheidungsbaumes vorhergesagt werden können. Dazu existieren verschiedene Möglichkeiten. Man verwendet neue, vom Entscheidungsbaum noch nicht gesehene Datensätze, um die Fehlerraten vorherzusagen. Die Fehlerraten können dann mit den Fehlerraten der Trainingsmenge verglichen werden. Dies hat bei geringen Datenmengen natürlich den Nachteil, dass zur Erstellung des Baumes weniger Daten verwendet werden können.

Quinlan hat einen Pruning–Ansatz entwickelt, bei dem nur die Trainingsdatensätze, aus denen der Entscheidungsbaum erstellt wurde, verwendet werden. Dazu betrachtet er die N Datensätze, die von einem Blatt abgedeckt werden, als N–Versuche bzw. N–malige Wiederholung eines Versuches. Die Menge E der fehlerhaften Datensätze im Blatt können als Ereignis (= Fehler) betrachtet werden. Damit kann die stark vereinfachte Annahme getroffen werden, dass die Fehlerwahrscheinlichkeit binomialverteilt $B(n,p)$ ist. Für die vorhergesagte Fehlerwahrscheinlichkeit p gilt dann:

$$P(X = E) \quad = \quad \binom{N}{E} \cdot p^E \cdot (1-p)^{N-E} \tag{3.26}$$

Der Wert $P(X = E)$ entspricht der Wahrscheinlichkeit, dass im Blatt genau E Fehler auftreten.
Die vorhergesagte Wahrscheinlichkeit eines Teilbaumes ergibt sich aus der Summe der vorhergesagten Wahrscheinlichkeiten aller Blätter dieses Teilbaumes. Der Teilbaum wird durch ein Blatt ersetzt, wenn diese Summe größer ist als die vorhergesagte Fehlerwahrscheinlichkeit des Blattes.

3.3.6 Verwendung von Entscheidungbäumen

Die Entscheidungsbäume versuchen ebenfalls logische Regeln für die Klassifikation zu finden. Sie haben zudem den Vorteil, dass sie übersichtliche Baumstrukturen mit relativ leicht nachvollziehbaren und überschaubaren Pfaden erzeugen. Dies erlaubt eine zusätzliche Validierung der in Kapitel 4 ermittelten Regelsätze.

In diesem Kapitel wurden drei unterschiedliche Klassifikationsverfahren vorgestellt. Wir wollen sie dazu verwenden, die Klassifikationsgüte unseres Modells zu überprüfen. Daneben können die erzeugten Regeln der logischen Modelle mit denen des Implikationenmodells verglichen werden. Zusätzlich können wir die Ergebnisse der logischen Modelle mit den Ergebnissen der neuronalen Netze vergleichen, die deutliche Unterschiede in der Vorgehensweise zur Klassifikation aufweisen.

Kapitel 4

Entwicklung eines verbandstheoretischen Implikationenmodells

4.1 Grundidee/Motivation

Im folgenden Kapitel wird ein verbandstheoretisches Klassifikationsmodell entwickelt, das folgenden Ansprüchen genügen soll:

- Überschaubarkeit der Regelsätze
- Transparenz für den Anwender
- Erfassung vieler Datensätze mit den Regeln
- Ermittlung von Wahrscheinlichkeiten für verschiedene Klassen
- einfache Implementierung
- Übertragbarkeit auf andere Fragestellungen

Vorab stellt sich natürlich die Frage, welche Merkmale bzw. Merkmalskombinationen sinnvoll zur Klassifikation verwendet werden sollen. Die Datensätze enthalten umfangreiche Informationen in Form einer Vielzahl von Variablen. Ziel ist es, aus dieser Vielzahl von Variablen, diejenigen zu extrahieren, die für die Klassenzugehörigkeit eines Datensatzes verantwortlich sind. Häufig sind dies nicht nur einzelne Merkmale, sondern bestimmte Merkmalskombinationen. Wie können solche Merkmalskombinationen aus dem umfangreichen Datenmaterial gewonnen werden? Um Zusammenhänge und Strukturen in Daten zu beschreiben, bietet die formale Begriffsanalyse eine Vielzahl von Anwendungen, die sich vor allem durch ihre Transparenz für den Anwender auszeichnen. Daher wollen wir im Folgenden die Theorie der formalen Begriffsanalyse

zur Modellbildung verwenden. Mit Hilfe der Implikationentheorie wird ein signifikantes Regelwerk zur Klassifikation erzeugt.

Um der Überschaubarkeit des Regelwerks Rechnung zu tragen, sind weitere Arbeitsschritte notwendig. Signifikante Regeln müssen herausgearbeitet und die Vielzahl der gewonnenen Regeln muss geeignet zusammengefasst werden. Zur Signifikanzprüfung und Zusammenfassung der Regeln werden statistische Testverfahren verwendet. In einem weiteren Schritt wird das Regelwerk bewertet und mit einem Bayesschen Ansatz bedingte Wahrscheinlichkeiten für die Regelsätze ermittelt.

Im Anwendungsteil der vorliegenden Arbeit wird die Modellierung dann zur Klassifikation realer Bauspardarlehen und zur Ermittlung von Kreditausfallwahrscheinlichkeiten verwendet. Zur Validierung und Überprüfung der Generalisierungsfähigkeit der Modellierung, wird das Verfahren auf die Originaldaten einer weiteren Bausparkasse übertragen. Im Anschluss daran wird auf eine Verwendung der Ergebnisse im Rahmen von IRB–Ansätzen eingegangen. Abschließend werden zwei weitere bauspartechnische Fragestellungen anhand realer Kollektivdaten mit derselben Vorgehensweise untersucht.

Zur logischen Modellierung müssen die Daten in binärer Form vorliegen, d. h. rationale Merkmalsausprägungen, wie z. B. die Darlehenshöhe oder das Alter eines Sparers, müssen diskretisiert werden. Die Vorgehensweise zur Ermittlung binärer Daten wird im folgenden Abschnitt ausführlich erläutert.

4.2 Datenvorbereitung/Diskretisierung

4.2.1 Bisherige Diskretisierungsverfahren

Im vorigen Kapitel wurde bereits das „Cutpoint–Verfahren" und ein entropiebasierter Ansatz zur Diskretisierung vorgestellt. Einen Vergleich der beiden Ansätze, sowie weitere Verfahren zur Erzeugung logischer Daten, findet man in [MT04]. Dort wurden Datensätze mit verschiedenen Verfahren diskretisiert und die Klassifikationsergebnisse, die mit vier unterschiedlichen Modellen erzeugt wurden, miteinander verglichen. Die Untersuchung lieferte das Ergebnis, dass die Güte des jeweiligen Diskretisierungsverfahrens stark vom verwendeten Klassifikationsmodell abhängt.

4.2.2 Ansatz der inhaltlichen Diskretisierung

Im vorliegenden Datenmaterial sind eine Vielzahl „bauspartechnischer" Informationen enthalten, die sich größtenteils aus der Tarifgestaltung der Bausparkassen ergeben. Es bietet sich daher an, dieses Expertenwissen bei der Diskretisierung der Daten einfließen zu lassen. Eine eingehende Analyse der in den Daten vorhandenen rationalen

Merkmale bestärkte diese Vorgehensweise [EFM04].
Zur Modellierung unserer Fragestellung können beispielsweise die Merkmale Darlehenshöhe, Spardauer, Alter und Höhe der Wohnungsbauprämie (WoP) verwendet werden. Die Bedeutung der inhärenten bauspartechnischen Informationen, und daraus resultierend das Vorgehen zur Diskretisierung, wird am Merkmal Spardauer kurz dargestellt. Die Spardauer bezeichnet die Anzahl der Jahre vom Abschluss eines Bausparvertrages bis zur Zuteilung. Dabei ist der Bausparer frei in seinem Sparverhalten. Allerdings schlagen die Bausparkassen den Sparern in ihren Tarifwerken bestimmte Regelsparraten vor, um möglichst „effizient" zum Sparziel zu gelangen.
Tabelle 4.1 zeigt die notwendigen Tarifparameter zur Ermittlung der Spardauer. Die Regelsparrate ist in Promille pro Monat angegeben und bezieht sich auf die volle Bausparsumme. Die Mindestwartezeit in Monaten bezeichnet die Anzahl der Monate nach Abschluss, in denen der Bausparvertrag nicht zugeteilt werden darf. Die vorgeschla-

Tarifname	Anspargrad in %	Regelsparrate	Mindestwartezeit
Tarif 1	40	4	18
Tarif 2	50	7	18
Tarif 3	40	4	18
Tarif 4	50	4	60

Tabelle 4.1: Auszug aus den Tarifkonditionen der Ausgangsbausparkasse[1]

gene Regelsparrate ist für den Bausparer nicht verbindlich, dennoch gibt es einen erheblichen Anteil sogenannter Regelsparer, die sich an der Rate orientieren. Die Mindestwartezeit gibt eine untere Schranke für die Spardauer an.

Beispiel 4.1. *Ermittlung der Spardauer in Abhängigkeit der Regelsparrate*
Der Ansatz

$$Anspargrad \approx x \cdot Regelsparrate \cdot 12$$

liefert für den Tarif 1

$$\begin{aligned} 0.4 &\approx x \cdot 0.004 \cdot 12 \\ x &\approx 8.33 \quad (Jahre). \end{aligned}$$

Da in dieser Berechnung keine Zinszahlungen berücksichtigt sind, wird der Wert abgerundet und man erhält eine Spardauer von acht Jahren im Tarif 1. Mit demselben Verfahren ergibt sich eine Spardauer von sechs Jahren im Tarif 2, acht Jahren im Tarif 3 und zehn Jahren im Tarif 4.

[1] Die angegebenen Tarifparameter beziehen sich auf ältere Tarifgenerationen und wurden umbenannt. Neuere Tarife wurden in der Untersuchung nicht berücksichtigt.

Die Intervalle für die Diskretisierung orientieren sich an den so ermittelten Spardauern. Da die ermittelten Größen nur Richtwerte und keine feste Vorgabe an den Bausparer darstellen, werden die Intervalle für alle Tarife verwendet. Ein Bausparer in Tarif 1 oder 4 kann durchaus seine Sparphase innerhalb sechs Jahren, z. B. durch Soforteinzahlungen beenden. Als untere Schranke im Tarif 4 kann eine Mindestlaufzeit von fünf Jahren angegeben werden, die durch unsere Intervalle über alle Tarife nicht verletzt wird. Eine kurze Spardauer beträgt weniger als sechs Jahre, eine mittlere Spardauer liegt zwischen sechs und zehn Jahren, und eine Spardauer von über zehn Jahren bezeichnen wir im Folgenden als lang. Die Diskretisierung der weiteren rationalen Variablen wurde mit Expertenwissen auf ähnliche Weise durchgeführt und ist im Anwendungsteil dieser Arbeit dargestellt.
Zum Vergleich der Klassifikationsmodelle mit dem verbandstheoretischen Implikationenmodell wurde die inhaltliche Diskretisierung für das Modell von Truemper und die Entscheidungsbäume verwendet. Daneben wurden die Modelle mit ihrer ursprünglich implementierten Diskretisierungsmethode getestet. Eine ausführliche Darstellung dieser Ergebnisse findet sich ebenfalls im zweiten Teil dieser Arbeit.

4.3 Ermittlung der Merkmalskombinationen

In den folgenden Ausführungen wird das generelle Vorgehen zur Ermittlung von signifikanten Merkmalskombinationen dargestellt. Die Vorgehensweise wird allgemein und unabhängig von der zugrunde liegenden Problemstellung formuliert.

4.3.1 Problemformulierung und Lösungsansätze

Mit Hilfe der inhaltlichen Diskretisierung erhält man eine (m,n)–Matrix mit m Gegenständen und n binären Variablen. Die Menge aller Gegenstände bezeichnen wir mit G, die Menge aller binären Merkmale mit M. Dabei gilt $|G| = m$ und $|M| = n$. Anhand der Matrix sollen für die Klassifikation relevante Merkmale bzw. Merkmalskombinationen ermittelt werden. Triviale Merkmale, d. h. Merkmale die keinen Beitrag zur Klassifikation leisten, sollen ebenfalls erkannt werden bzw. im Regelwerk nicht mehr auftauchen. Da es sich um ein überwachtes Klassifikationsproblem handelt, ist das Zielattribut $w \notin M$ (nämlich die Klassenzugehörigkeit) bekannt, das unsere Ausgangsmatrix in zwei Teilmatrizen zerlegt. Zum einen die Teilmatrix G^+, die alle Gegenstände umfasst, die das Zielattribut w besitzen, und die Teilmatrix G^-, die alle Gegenstände enthält, die w gerade nicht besitzen. Es sollen Regeln formuliert werden, welche die strukturellen Zusammenhänge innerhalb beider Teilmatrizen erfassen und damit unsere Ausgangsvermutung bestätigen.

In [FT02] wird mit Hilfe der „Clash– " und „Agree–Bedingungen" eine trennende DNF–Formel $D = D_1 \vee D_2 \vee \ldots \vee D_k$ ermittelt, welche die Datensätze gemäß ihrer

Klassenzugehörigkeit separiert. Zur Bewertung ungesehener Testfälle wird die Anzahl der erfüllten Formeln pro Datensatz gezählt. Allerdings werden bei der Formulierung der Bedingungen alle Variablen verwendet, ohne deren Beitrag zur Klassifikation zu untersuchen. Dies widerspricht unserer Zielvorstellung, dass signifikante Strukturen herausgearbeitet werden sollen, um ein überschaubares Regelwerk zu erhalten. Auch für triviale Merkmale bzw. Merkmalskombinationen werden Bedingungen formuliert. Wir suchen daher nach einem Verfahren, dass uns die innere Struktur der beiden Matrizen G^+ und G^-, die wir im Folgenden auch als Mengen bezeichnen wollen, herausarbeitet. Wir werden einen verbandstheoretischen Ansatz zur Ermittlung der Merkmalskombinationen wählen, da die formale Begriffsanalyse geeignete Verfahren zur strukturellen Untersuchung von Daten liefert.

In [GK00, GK01] wird ein Ansatz vorgeschlagen, bei dem mit Hilfe von Implikationen Hypothesen zur Klassifizierung erstellt werden. Ganter und Kuznetsov gehen dabei ebenfalls von zwei Grundmengen aus, die positiven Beispiele (G^+), welche das Zielattribut w besitzen und die negativen Beispiele (G^-), die w nicht besitzen. Diese Hypothesen sollen Objekte mit unbekanntem Zielattribut klassifizieren. Grundannahme bei der Formulierung von positiven Hypothesen ist, dass bestimmte Mengen von Attributen existieren, die nur von positiven Beispielen erfüllt werden, aber von keinem negativen Beispiel. Dasselbe gilt natürlich für die Formulierung von negativen Hypothesen. Positive und negative Hypothesen sind folgendermaßen definiert:

Definition 4.1. *[GK00] Eine positive Hypothese im Hinblick auf das Zielattribut w, ist ein Gegenstandsinhalt des Kontextes $\mathbb{K}^+ = (G^+, M, I)$, der nicht Gegenstandsinhalt g' eines negativen Beispiels $g \in G^-$ ist. Ein Objekt $g \in G^+$ besitzt eine positive Hypothese, wenn*

$$g' \quad := \quad \{m \in M \mid (g,m) \in I\}$$

eine positive Hypothese enthält. Analoges gilt für negative Hypothesen.

Mit Hilfe der Implikationentheorie werden Hypothesen in Form von Gegenstandsinhalten formuliert, die zur Klassifikation unbestimmter Objekte verwendet werden. Ein Objekt besitzt dann das Zielattribut w, falls es eine positive Hypothese erfüllt, und es besitzt das Zielattribut nicht, falls es eine negative Hypothese erfüllt. Ein unbestimmtes Objekt wird positiv bewertet, wenn es eine positive, aber keine negative Hypothese erfüllt. Die negative Zuordnung eines unbestimmten Objektes verläuft ebenso.

Diese Vorgehensweise erscheint für unsere Problemstellung jedoch nicht praktikabel. Allein die Grundannahme, dass Hypothesen in Form von Gegenstandsinhalten existieren, die ausschließlich in einer der beiden Mengen G^+ oder G^- auftreten, bereitet bei dem vorliegenden umfangreichen Datenmaterial Probleme. Ein kurzes Beispiel soll dies verdeutlichen.

Beispiel 4.2. *In den untersuchten Mengen G^+ und G^- ist die binäre Variable „Weiteres Darlehenskonto in Zahlungsschwierigkeiten? Ja oder Nein" folgendermaßen verteilt:*

Variable vorhanden?	*Menge G^+ (in %)*	*Menge G^- (in %)*
Ja	*32.83*	*0.10*
Nein	*67.17*	*99.90*

Hat ein Darlehensnehmer bereits ein Darlehenskonto, das sich in Zahlungsverzug befindet, so ist davon auszugehen, dass das Darlehenskonto ebenfalls in Zahlungsschwierigkeiten geraten wird.

Allerdings existieren im vorliegenden Datenmaterial auch endgetilgte Konten ($g \in G^-$), die ein weiteres Darlehenskonto in Zahlungsverzug besitzen. Daraus ließe sich mit dem eben vorgestellten Ansatz keine positive Hypothese formulieren, obwohl eine solche Hypothese sinnvoll wäre.

Eine weitere Schwachstelle liegt in der Bewertung der unbestimmten Objekte. Es existieren sicherlich Konten, die sowohl positive als auch negative Hypothesen erfüllen. Daneben drängt sich dem Leser die Frage auf, ob ein Konto das mehrere positive Hypothesen erfüllt, genauso zu bewerten ist, wie ein Konto das nur eine positive Hypothese erfüllt. Hier erscheint eine differenzierte Bewertung der erfüllten Hypothesen angebracht.
Um die angeführten Nachteile zu beseitigen, werden wir in unserer Modellierung die Annahme der Eindeutigkeit der Hypothesen verwerfen und die kombinierbaren ermittelten Merkmalskombinationen empirisch bewerten. Allerdings wollen wir die Regeln ebenfalls auf Basis der Implikationentheorie formulieren. Mit Hilfe des Zielattributes $w \notin M$ wollen wir die Datenmenge in zwei Teilkontexte $\mathbb{K}^+ := (G^+, M, I^+)$ sowie $\mathbb{K}^- := (G^+, M, I^-)$ zerlegen.

Um Merkmalszusammenhänge zu erarbeiten, bietet die formale Begriffsanalyse verschiedene Methoden. Zum einen können alle Begriffsinhalte ermittelt und daraus ein Liniendiagramm erstellt werden. Damit erhält man einen Überblick über die Struktur des Kontextes. Allerdings kann die Anzahl der Begriffe exponentiell mit der Größe des Kontextes wachsen [Lin99]. So kann ein Kontext (G, M, I) mit $|G| = m$ und $|M| = n$, maximal bis zu 2^l Begriffe besitzen, mit $l = min(m, n)$.
Um die hohe Anzahl von Begriffen zu reduzieren wird in [Stu02] nach häufigen Begriffen gesucht. Dazu werden für einen Kontext die sogenannten Träger („support") berechnet.

Definition 4.2. *[Stu02] Der Support einer Merkmalsmenge $X \subseteq M$ für einen Kontext (G, M, I) ist gegeben durch:*

$$supp(X) = \frac{|X'|}{|G|}$$

Damit können die Begriffsinhalte zusätzlich nach ihrer relativen Häufigkeit des Auftretens bewertet werden. Diese Art der Analyse erlaubt es, Kontexte bzw. Liniendia-

gramme zu erstellen, die nur die häufigsten Begriffsinhalte enthalten. Diese Begriffsverbände werden auch als Eisbergverbände bezeichnet. Eine Einführung in die Theorie der Eisbergverbände findet sich ebenfalls in [Stu02].
Unsere empirische Regelformulierung erlaubt im Gegensatz zu [GK00], dass eine Regel sowohl in der Menge der positiven als auch in der Menge der negativen Beispiele auftritt. Ein häufiger Begriff im Kontext $\mathbb{K}^+$ kann äquivalent als häufiger Begriff in $\mathbb{K}^-$ vorhanden sein. Im Hinblick auf unsere Zielformulierung wurden dadurch keine signifikanten Strukturen für die Mengen G^+ und G^- erschlossen. Daher müssen bei der Regelformulierung beide Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ berücksichtigt, und möglicherweise auch eine Vielzahl von Begriffen untersucht werden. Die Verwendung von Eisbergverbänden für unsere Problemstellung erscheint daher ebenfalls nicht praktikabel.

Bei einer hohen Anzahl von Gegenständen und einer relativ kleinen Menge von Merkmalen bietet es sich an, den Kontext anhand der Implikationen zwischen den Merkmalen zu erschließen [GW96]. Dies hat den Vorteil, dass die Anzahl der Implikationen deutlich geringer als die Anzahl der Begriffsinhalte ist, dadurch aber keine Informationen des Kontextes verloren gehen. Die Formulierung von Regeln mit Hilfe von Implikationen weist Ähnlichkeiten zur Definition 4.1 von Ganter und Kuznetsov auf, wie folgendes Lemma zeigt.

Lemma 4.1. *Seien $A, B \subseteq M$. Die Implikation $A \to B$ gilt genau dann in einem Kontext $\mathbb{K} = (G, M, I)$, wenn $A \to B$ für das System $g' := \{m \in M \mid (g,m) \in I\}$ der Gegenstandsinhalte gilt, also von jedem Gegenstandsinhalt respektiert wird.*

Beweis. Gelte $A \to B$ im Kontext $\mathbb{K} = (G, M, I)$ und sei $g \in G$. Es folgt: $A \subseteq g' \Rightarrow g'' \subseteq A' \Rightarrow g' = g''' \supseteq A'' \supseteq B$.
Gilt umgekehrt für jeden Gegenstandsinhalt g' die Beziehung $A \subseteq g' \Rightarrow B \subseteq g'$, so ergibt sich für alle $h \in G$: $h \in A' \Rightarrow A \subseteq h' \Rightarrow B \subseteq h' \Rightarrow h \in B'$. Damit ist $A' \subseteq B' \Rightarrow B \subseteq A''$ und $A \to B$ gilt in $\mathbb{K}$. □

Daraus folgt, dass $A \to B$ genau dann im Kontext $\mathbb{K} = (G, M, I)$ gilt, wenn $A \to B$ von jedem Begriffsinhalt respektiert wird.

Beweis. Wird $A \to B$ von jedem Begriffsinhalt respektiert, so auch von jedem Gegenstandsinhalt. Daraus folgt nach dem eben Bewiesenen $A \to B$ gilt in $\mathbb{K}$.
Gilt $A \to B$ in $\mathbb{K}$ und sei $T \subseteq M$ Begriffsinhalt, d. h. $T = T''$. Wir erhalten $A \subseteq T \Rightarrow A'' \subseteq T'' \Rightarrow B \subseteq T'' = T$. T respektiert also $A \to B$. □

Die mit Hilfe der Implikationentheorie formulierten Regeln respektieren also alle Begriffsinhalte der Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$. Ziel ist es, eine handliche Menge von Implikationen der Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ zu finden, aus denen anschließend Regeln zur Klassifikation erzeugt werden können.

4.3.2 Ermittlung der Stammbasis der Implikationen

Zur Ermittlung einer Stammbasis von Implikationen verwenden wir den Ganter–Algorithmus, der auch als Next–Closure–Algorithmus bezeichnet wird. Er ermittelt sukzessive alle Begriffs- und Pseudoinhalte und basiert auf einer totalen Ordnung auf den Begriffen. Ausgehend vom kleinsten Begriff wird der jeweilige Nachfolger ermittelt. Dazu definieren wir eine lexikographische Ordnung < auf der Menge aller Teilmengen von G, die auf einer lexikographischen Ordnung der Mengen G und M aufbaut: $M = \{m_1 < m_2 \ldots < m_n\}$ und analog für G.

Definition 4.3. Lexikographische Ordnung
Von zwei verschiedenen Mengen $A, B \subseteq M$ heißt A lexikographisch kleiner genau dann, wenn das kleinste Element, in dem sich A und B unterscheiden, in B enthalten ist. Formal:

$$\begin{aligned} A < B &:\Longleftrightarrow A <_i B \\ &:\Longleftrightarrow \exists_{i \in B \setminus A}\, A \cap \{1,2,\ldots,i-1\} = B \cap \{1,2,\ldots,i-1\} \end{aligned}$$

Zusätzlich ist die Operation $\oplus$, die eine neue Merkmalsmenge bestimmt, folgendermaßen definiert:

Definition 4.4. *Für $A \subseteq M$ sei:*

$$A \oplus i := ((A \cap \{1,2,\ldots,i-1\}) \cup \{i\})''$$

Für $A <_i B$ und $A \oplus i$ können folgende Aussagen verifiziert werden:

(1) $A < B :\Longleftrightarrow A <_i B$ für ein $i \in M$.
(2) $A <_i B$ und $A <_j C$ mit $i < j \Rightarrow C <_i B$.
(3) $i \notin A \Rightarrow A < A \oplus i$.
(4) $A <_i B$ und B Begriffsinhalt $\Rightarrow A \oplus i \subseteq B$, d. h. $A \oplus i \leq B$.
(5) $A <_i B$ und B Begriffsinhalt $\Rightarrow A <_i A \oplus i$.

Mit Hilfe der Operation $\oplus$ lässt sich zu einem Begriff der lexikographisch nächste Begriff bestimmen, wie der folgende Satz zeigt.

Satz 4.1. *[GW96] Der kleinste Begriffsinhalt, der bzgl. der lexikographischen Ordnung größer ist als eine gegebene Menge $A \subset M$, ist*

$$A \oplus i,$$

wobei i das größte Element von M ist mit $A <_i A \oplus i$.

Beweis. Sei A^+ der kleinste Inhalt nach A bzgl. der lexikographischen Ordnung. Wegen $A < A^+$ ist $A <_i A^+$ für ein $i \in M$ nach (1) und damit $A <_i A \oplus i$ nach (5). Nach (4) folgt $A \oplus i \leq A^+$, also $A \oplus i = A^+$ wegen $A < A \oplus i$. Dass i das größte Element ist mit $A <_i A \oplus i$, ergibt sich aus (2), denn $A <_j A \oplus j$ mit $j \neq i$ hat wegen $A \oplus i = A^+ < A \oplus j$ nach (2) $j < i$ zur Folge. □

Jetzt wollen wir uns mit Hilfe eines geeigneten Hüllensystems mit der lexikographischen Konstruktion von Begriffs- und Pseudoinhalten befassen. Als Grundlage benötigen wir folgenden Hilfssatz:

Hilfssatz 4.2. *[GW96] Sind P und Q Begriffs- oder Pseudoinhalte mit $P \neq Q, P \not\subset Q$ und $Q \not\subset P$, so ist $P \cap Q$ ein Begriffsinhalt.*

Beweis. Sowohl P als auch Q, und damit auch $P \cap Q$, respektieren alle Implikationen in $\mathcal{L}$ mit Ausnahme von $P \to P''$ und $Q \to Q''$. Ist $P \neq P \cap Q \neq Q$, so respektiert $P \cap Q$ auch diese Implikationen, ist also ein Inhalt. □

Als unmittelbare Folge des Hilfssatzes 4.2 erhält man:

Hilfssatz 4.3. *[GW96] Die Menge aller Teilmengen von M, die Begriffs- oder Pseudoinhalte eines Kontextes (G,M,I) sind, ist ein Hüllensystem.* □

Der Hüllenoperator zu diesem Hüllensystem entsteht durch Modifikation aus dem Operator $\mathcal{L}$. Ausgehend von einer Menge $X \in M$ bilden wir sukzessive

$$X^{\mathcal{L}^*} := X \cup \bigcup \{B \mid A \to B \in \mathcal{L}, A \subseteq X, A \neq X\}$$

$$X^{\mathcal{L}^*\mathcal{L}^*} := X^{\mathcal{L}^*} \cup \bigcup \{B \mid A \to B \in \mathcal{L}, A \subseteq X^{\mathcal{L}^*}, A \neq X^{\mathcal{L}^*}\}$$

und so fort, bis schließlich eine Menge $\mathcal{L}^*(X)$ mit $\mathcal{L}^*(X) = \mathcal{L}^*(X)^{\mathcal{L}^*}$ erreicht ist. Diese ist dann der gesuchte Pseudo- oder Begriffsinhalt.
Der Satz 4.1 ist der Kern von Ganters Algorithmus und zeigt, wie der gesuchte Begriffsinhalt zu finden ist. Das Vorgehen des Algorithmus kann auch folgendermaßen beschrieben werden:

1. Die Menge $\mathcal{L}$ aller Implikationen wird auf die leere Menge gesetzt.

2. Der lexikographisch kleinste Begriffs- oder Pseudoinhalt ist $\emptyset$.

3. Ist A als Begriffs- oder Pseudoinhalt bestimmt, so findet man den lexikographisch nächsten Begriffs- oder Pseudoinhalt, indem man alle Merkmale $i \in M \setminus A$ prüft, beginnend mit dem größten, und dann in absteigender Reihenfolge, bis erstmals $A <_i \mathcal{L}^*(A \oplus i)$ gilt.

   $\mathcal{L}^*(A \oplus i)$ ist dann der nächste Begriffsinhalt.

4. Gilt $\mathcal{L}^*(A \oplus i) = (\mathcal{L}^*(A \oplus i))''$, dann ist $\mathcal{L}^*(A \oplus i)$ ein Begriffsinhalt. Ansonsten ist er Pseudoinhalt und die Implikation $\mathcal{L}^*(A \oplus i) \to (\mathcal{L}^*(A \oplus i))''$ wird zu $\mathcal{L}$ hinzugefügt.

5. Wenn $\mathcal{L}^*(A \oplus i) = M$, dann Ende, sonst $A \leftarrow \mathcal{L}^*(A \oplus i)$ und weiter bei 3.

Die Vorgehensweise des Algorithmus soll an einem kurzen Beispiel mit Hilfe einer Tabelle dargestellt werden.

Beispiel 4.3. *Die Implikationen und Begriffsinhalte des Beispielkontextes sollen ermittelt werden. Dabei bezeichnet B.I. einen Begriffsinhalt, P.I. einen Pseudoinhalt und damit eine Implikation.*

	a	b	c	d
1		X		
2	X	X		
3			X	X
4			X	

A	i	$A \oplus i$	$\mathcal{L}^*(A \oplus i)$	$A <_i \mathcal{L}^*(A \oplus i)$?	$(\mathcal{L}^*(A \oplus i))''$	$\mathcal{L}$	Inhalte
$\emptyset$	d	d	d	$\emptyset <_d$ d Ja	cd	$d \to c$	d=P.I.
d	d	c	c	$d <_c c$ Ja	c	-	c=B.I.
c	d	cd	cd	$d <_d cd$ Ja	cd	-	cd=B.I.
cd	b[2]	b	b	$cd <_b b$ Ja	b	-	b=B.I.
b	d	bd	bdc	$b <_d bdc$ Nein wg.c			
b	c	bc	bc	$b <_c bc$ Ja	abcd	$bc \to ad$	bc=P.I.
bc	d	bcd	abcd	$bc <_d abcd$ Nein wg.a	-	-	-
bc	a	a	a	$bc <_a a$ Ja	ab	$a \to b$	a=P.I.
a	d	ad	abcd	$a <_d abdc$ Nein wg.c	-	-	-
a	c	ac	abcd	$a <_c abdc$ Nein wg.b	-	-	-
a	b	ab	ab	$a <_b ab$ Ja	ab	-	ab=B.I.
ab	d	abd	abcd	$ab <_d abcd$ Nein wg.c	-	-	-
ab	c	abc	abcd	$ab <_c abcd$ Ja	abcd	-	abcd=B.I.

Tabelle 4.2: Durchführung des Next–Closure–Algorithmus für den Beispielkontext

Der Algorithmus liefert folgende Stammbasis der Implikationen:

$$\mathcal{L} = \{d \to c, bc \to ad, a \to b\}$$

Zusätzlich werden die Begriffsinhalte ausgegeben, die im Folgenden nur durch ihre Merkmale angegeben werden.

$$\mathcal{B} = \{(\emptyset), (c), (cd), (b), (ab), (abcd)\}$$

Der Algorithmus Next–Closure berechnet alle Begriffs- und Pseudoinhalte eines Kontextes mit der Komplexität $O(|M|^2 \cdot |G| \cdot |\mathcal{B}(G,M,I)|)$ [Lin99]. Die Ableitungsoperation $''$ mit ihrer Komplexität $O(|M| \cdot |G|)$ muss maximal $|M|$-mal ausgeführt werden. Dabei bezeichnet $|\mathcal{B}(G,M,I)|$ die Anzahl der Begriffe, die exponentiell mit der Größe

[2] Die Merkmale d und c müssen nicht mehr untersucht werden, da sie im Begriffsinhalt (cd) selbst liegen. Diese Schritte werden im Folgenden nicht mehr aufgeführt.

des Verbandes wachsen kann. In [Lin99] wird unter anderem eine Komplexitätsreduktion des Algorithmus vorgestellt. Diese bezieht sich auf Kontexte, die aus einer Ergänzung eines bestehenden Kontextes entstehen. Der Kontext kann unter Verwendung des Verbandes des ursprünglichen Kontextes berechnet werden. Dadurch wird eine vollständige Neuberechnung des Verbandes vermieden und die gewünschte Komplexitätsreduktion erreicht.
In [KO02] werden weitere Algorithmen zur Ermittlung von Begriffs- und Pseudoinhalten vorgestellt und Laufzeitvergleiche durchgeführt. Dabei zeigte sich, dass die Laufzeiten der verschiedenen Algorithmen sehr stark von der Größe des zu untersuchenden Kontextes abhängen. Aufgrund der Ergebnisse von [KO02] und der bestehenden Implementierung in ConImp[3] wurde zur Ermittlung der Stammbasis der Algorithmus Next–Closure gewählt.

4.4 Auswahl der ermittelten Merkmalskombinationen

Mit Hilfe des Next–Closure–Algorithmus sind wir jetzt in der Lage aus den Kontexten $\mathbb{K}^+$ und $\mathbb{K}^-$ die Stammbasis der Implikationen zu ermitteln. Alle Gegenstände des jeweiligen Kontextes respektieren diese Implikationen. Allerdings ist die Anzahl der Implikationen zur Ermittlung eines überschaubaren Regelwerkes immer noch sehr hoch. Unser Ziel ist es daher, die für die Klassifikation relevanten Implikationen aus der Stammbasis herauszuarbeiten. Dazu bedarf es allerdings einiger Vorarbeiten.

Um die Anzahl der Implikationen der Stammbasis zu reduzieren, werden zum einen nur Implikationen ausgewählt, die nicht trivialerweise von den Trainingsbeispielen erfüllt werden. Für unsere empirische Untersuchung wollen wir ausschließlich Implikationen $A \rightarrow B$ mit $A, B \subseteq M$ mit erfüllter Prämisse betrachten.

Eine weitere Reduzierung der Implikationen wird dadurch erreicht, dass der Anteil der Trainingsbeispiele, welche die Implikation echt erfüllen, einen bestimmten Schwellwert $\delta \in \mathbb{R}$ überschreiten muss. Es existieren Implikationen in der Stammbasis, die nur von wenigen Objekten echt erfüllt werden. Ziel unseres Modells ist es jedoch, grundlegende Strukturen bzw. signifikante Implikationen herauszuarbeiten, mit deren Hilfe anschließend viele Objekte richtig klassifiziert werden können. Der Schwellwert δ wird in unserem Modell auf 10 % gesetzt, d. h. Implikationen für die gilt:

$$\frac{|\hat{g}|}{|G|} < \delta \quad \text{mit} \quad \hat{g} = \{g' \,|g \in G \text{ respektiert } A \rightarrow B\}$$

[3] Das Programm kann kostenlos unter `http://www.mathematik.tu-darmstadt.de/~burmeister/` heruntergeladen werden.

werden in unserer Untersuchung vernachlässigt. Diese Vorgehensweise kann intuitiv damit begründet werden, dass durch diese Implikationen Einzelfälle abgebildet werden, die bei der Klassifikation einer Testmenge zu einem Overfitting–Effekt führen.
Implikationen, die von vielen Objekten echt erfüllt werden, liegen im zugehörigen Begriffsverband $\underline{\mathcal{B}}(G,M,I)$ in der Nähe des Einselements, auf dem alle Gegenstände liegen. Dementsprechend liegen auf dem Nullelement alle Merkmale. Je näher ein Merkmal $A \subseteq M$ am Einselement in $\underline{\mathcal{B}}(G,M,I)$ liegt, umso mehr Gegenstände $g \in G$ besitzen das Merkmal. Ebenso gilt, je näher ein Gegenstand $g \in G$ in $\underline{\mathcal{B}}(G,M,I)$ am Nullelement liegt, umso mehr Merkmale $m \in M$ besitzt er. Häufige Implikationen liegen im Hasse–Diagramm also näher am Einselement und spiegeln daher den grundlegenden Aufbau von $\underline{\mathcal{B}}(G,M,I)$ wider.

4.5 Ermittlung von signifikanten Implikationen für die Mengen G^+ und G^-

4.5.1 Auswahl des Testverfahrens

Nach Auswahl von häufigen, echt erfüllten Implikationen stellt sich die Frage, ob die Implikationen der Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ auch signifikant für die Mengen G^+ und G^- sind. Eine Implikation des Kontextes $\mathbb{K}^+$ kann genauso häufig in der Menge G^+, wie in der Menge G^- auftreten. Analoges gilt für Implikationen des Kontextes $\mathbb{K}^-$. Es muss eine Prüfstatistik gewählt werden, die in der Lage ist zu testen, ob eine Implikation des Kontextes $\mathbb{K}^+$ signifikant häufiger in der Menge G^+ als in der Menge G^- auftritt. Die Statistik stellt hier eine Vielzahl an Hypothesentests zur Verfügung. Eine Übersicht der Testverfahren findet sich z. B. in [BLK06].

Die vorliegende Fragestellung kann auch folgendermaßen formuliert werden: „Ist das Auftreten von Implikation *i* unabhängig von der Menge aus der das Beispiel stammt ?" Damit haben wir eine erste statistische Unabhängigkeitshypothese formuliert.
Zum Testen von Unabhängigkeitshypothesen existieren verschiedene Prüfstatistiken. In Kapitel 2 wurde der exakte Test von Fisher vorgestellt, der den Vorteil hat, dass die Prüfgröße nicht durch eine Verteilung approximiert werden muss, sondern direkt ermittelt werden kann. Allerdings ist der Test eher für kleinere Stichprobenumfänge geeignet, da sich hier die Verteilung der Testgröße ohne allzu großen Aufwand berechnen lässt. Bei Vorlage größerer Stichprobenumfänge, wie in unserem Falle, erscheint dieses Testverfahren ungeeignet.
Daher entscheiden wir uns zum Überprüfen der Unabhängigkeitshypothese für den χ^2–Unabhängigkeitstest. Der χ^2–Unabhängigkeitstest überprüft, ob zwei oder auch mehrere Variablen stochastisch unabhängig sind. Damit die benutzte χ^2–Verteilungsapproximation hinreichend genau ist, sollte $n_{ij} \geq 5$ für alle $i, j = 1,2$ gelten [CL03,

BB96]. Diese Voraussetzung ist in unserer Untersuchung stets erfüllt.

4.5.2 Formulierung von Hypothesen und der Prüfstatistik

Bei der Formulierung der Hypothesen und der Prüfstatistik beschränken wir uns in den folgenden Ausführungen auf den vorliegenden Fall $k = l = 2$. Für eine zweidimensionale Zufallsvariable (X,Y) soll überprüft werden, ob ihre Komponenten unabhängig sind. Dazu zerlegen wir den Wertebereich von X und Y in zwei Mengen X_1, X_2 und Y_1, Y_2. Für $i, j = 1,2$ sei:

$$\begin{aligned} p_{ij} &= P(X \in X_i, Y \in Y_j) \quad \text{und} \\ p_{i\bullet} &= \sum_{j=1}^{2} p_{ij} = P(X \in X_i) \quad \text{sowie} \\ p_{\bullet j} &= \sum_{i=1}^{2} p_{ij} = P(Y \in Y_j). \end{aligned}$$

Sind die Zufallsvariablen stochastisch unabhängig, so müsste gelten:

$$P(X \in X_i, Y \in Y_j) = P(X \in X_i) \cdot P(Y \in Y_i),$$

also $p_{ij} = p_{i\bullet} \cdot p_{\bullet j}$. Daraus ergeben sich die zu testenden Hypothesen:

$$\begin{aligned} H_0 &:= p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \quad \text{für alle Paare } (i,j) \\ H_1 &:= p_{ij} \neq p_{i\bullet} \cdot p_{\bullet j} \quad \text{für mindestens ein Paar } (i,j). \end{aligned}$$

Als Signifikanzniveau wird $\alpha = 0.01$ gewählt. Die Prüfgröße T wird folgendermaßen ermittelt:

$$T = n \cdot \frac{(n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2}{n_{1\bullet} \cdot n_{2\bullet} \cdot n_{\bullet 1} \cdot n_{\bullet 2}}$$

Die Nullhypothese H_0 wird abgelehnt wenn $T > \chi^2_{(1;1-\alpha)}$ ist.

4.5.3 Testdurchführung

Um eine möglichst hohe Signifikanz der Regeln zu gewährleisten wurde α, dass dem Risiko einer Fehlentscheidung entspricht, sehr gering gewählt. Die Durchführung des Tests wird beispielhaft an der Implikation I_1 dargestellt.

Beispiel 4.4. *Die Implikation „Selbständig → Tarif 1/Tarif 3" soll bezüglich ihrer Signifikanz des Auftretens in den beiden Mengen G^+ und G^- getestet werden.*
Es wird also die Nullhypothese
H_0 := Das Vorhandensein der Implikation I_1 ist unabhängig davon, aus welcher Menge

das Konto stammt
gegen die Alternativhypothese
H_1 := Das Vorhandensein der Implikation I_1 ist abhängig davon, aus welcher Menge das Konto stammt
getestet. Die empirische Auswertung der Daten lieferte die Kontingenztabelle 4.3.

	ausg. Konten	endg. Konten	$\sum$
I_1	160	1.142	1.302
$\neg I_1$	1.205	26.734	27.939
$\sum$	1.365	27.876	29.241

Tabelle 4.3: Kontingenztabelle der Implikation I_1

Der kritische Bereich bei einem Signifikanzniveau $\alpha = 0.01$ ist näherungsweise-$\chi^2_{(k-1)(l-1)}$–verteilt und beträgt 6.64.
Für unser Beispiel wird H_0 abgelehnt, wenn für die Prüfgröße T gilt:

$$T = n \cdot \frac{(n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2}{n_{1\bullet} \cdot n_{2\bullet} \cdot n_{\bullet 1} \cdot n_{\bullet 2}} > 6.64$$

$$T = 29.241 \cdot \frac{(160 \cdot 26.734 - 1.142 \cdot 1.205)^2}{1.302 \cdot 27.939 \cdot 1365 \cdot 27876} \approx 177.83 > 6.64$$

$\Rightarrow$ *Die Unabhängigkeitshypothese H_0 wird auf dem Signifikanzniveau α verworfen, die Alternativhypothese H_1 wird angenommen.*

Der χ^2–Unabhängigkeitstest wurde für alle in Abschnitt 4.4 ermittelten Implikationen der Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ durchgeführt. Die Implikationen, deren χ^2–Wert unterhalb des kritischen Bereiches lag, wurden verworfen, da sie nicht dazu dienen signifikante Strukturen aus den Mengen G^+ und G^- herauszuarbeiten. Damit erhält man zwei Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ von Implikationen, die wir im Folgenden auch als Regeln bezeichnen wollen. Allerdings müssen sie noch zu einem handlichen Regelwerk transformiert werden.

4.6 Erstellung eines kompakten Regelwerkes

4.6.1 Ausgangssituation

Aus den oben erwähnten Schritten erhält man zwei Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ mit bereinigten und signifikanten Implikationen r_i, $i = 1,2,\ldots,n$ und $\hat{r}_j$, $j = 1,2,\ldots,m$ mit zugehörigen χ^2–Werten für die Mengen G^+ und G^-. Da es sich bei den Regeln stets um Implikationen mit erfüllter Prämisse handelt, können sie auch in Form von Konjunktionen dargestellt werden, z. B. $r_n = x_i \wedge x_j \wedge \cdots \wedge x_k$ oder $\hat{r}_m = x_o \wedge x_p \wedge \cdots \wedge x_r$. Es wird jetzt nach einer Bewertung der Regeln in Tabelle 4.4 gesucht, die in der Lage ist, die

$\mathcal{L}_1$	$\chi^2_{r_n}$	$\mathcal{L}_2$	$\chi^2_{\hat{r}_m}$
r_1	$\chi^2_{r_1}$	$\hat{r}_1$	$\chi^2_{\hat{r}_1}$
r_2	$\chi^2_{r_2}$	$\hat{r}_2$	$\chi^2_{\hat{r}_2}$
...	...	...	...
...	...	...	...
...	...	...	...
r_n	$\chi^2_{r_n}$	$\hat{r}_m$	$\chi^2_{\hat{r}_m}$

Tabelle 4.4: Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ mit zugehörigen Regeln r_n und $\hat{r}_m$, sowie den jeweiligen χ^2–Werten.

Trainingsbeispiele richtig zu klassifizieren. Weiterhin soll durch die Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ eine maximale Anzahl von Datensätzen überdeckt werden. Allerdings kann ein Datensatz k_i mehrere Regeln aus $\mathcal{L}_1$ oder $\mathcal{L}_2$ erfüllen. Möglich sind auch Kombinationen beider Mengen, wie folgendes Beispiel verdeutlichen soll.

Beispiel 4.5. *Datensatz k_i erfüllt die Regeln r_1, r_5 aus $\mathcal{L}_1$. Ist k_i besser oder schlechter zu bewerten als ein Datensatz k_j der nur r_5 erfüllt?*

Datensatz k_l erfüllt die Regeln r_4, r_5 aus $\mathcal{L}_1$ und $\hat{r}_3, \hat{r}_5$ aus $\mathcal{L}_2$. Ist k_l schlechter oder besser zu bewerten, da er Regeln aus beiden Mengen erfüllt.

Im Modell von Truemper wird zur Lösung des Problems die Anzahl der erfüllten Formeln gezählt und addiert. Dabei liegen allerdings die Annahmen zu Grunde, dass alle Regeln gleichwertig einzustufen sind, und dass ein Datensatz der mehr Regeln erfüllt auch automatisch „besser" eingestuft wird, als ein Datensatz mit weniger erfüllten Regeln. Diese Annahmen wollen wir aber durch eine empirische Bewertung aller Mengen von Teilmengen aus $\mathcal{L}_1$ und $\mathcal{L}_2$ gerade erst nachweisen. Daher muss die Potenzmenge aus $\mathcal{L}_1$ und $\mathcal{L}_2$ in Betracht gezogen und bewertet werden. Die Gesamtanzahl aller Regeln beider Mengen beträgt $n+m=k$ Regeln[4]. Daher liefert die kombinatorische Teilmengenbildung $|\mathcal{P}(k)|=2^k$ mögliche Teilmengen. Allerdings beträgt die Anzahl der möglichen Kombinationen z. B. bei $n=m=5$ schon $2^{10}=1.024$ Teilmengen, bei $n=m=10$ beträgt die Anzahl bereits $2^{20}=1.048.576$ Teilmengen.
Die Potenzmenge $\mathcal{P}(k)$ muss mit dem zur Verfügung stehenden Datenmaterial empirisch bewertet werden.
Allerdings treten hier zwei Probleme unterschiedlicher Art auf. Zum einen bereitet die Enummerierung von $\mathcal{P}(k)$ rechentechnische Schwierigkeiten, da die Laufzeit exponentiell mit der Anzahl der Regeln k wächst. Zum anderen bereitet die empirische Bewertung einer hohen Anzahl von Teilmengen Probleme. Trotz einer vergleichsweisen großen Anzahl von Trainings- und Testdaten ist nicht gewährleistet, dass alle Teilmengen mit Datensätzen bzw. mit einer ausreichenden Anzahl von Datensätzen belegt

[4]Die Anzahl der Regeln in beiden Mengen muss nicht identisch sein.

werden können. Dadurch ist keine statistisch abgesicherte Bewertung und daraus resultierend auch keine Klassifikation der Datensätze möglich.

4.6.2 Zusammenfassung von Regeln

Zur Lösung beider Probleme bietet es sich an, die Regeln der jeweilige Menge G^+ und G^- zusammenzufassen. Dies wirft allerdings die Frage auf, anhand welcher Gesichtspunkte die Zusammenfassung geschehen soll. Es soll ein Verfahren entwickelt werden, dass es stets erlaubt die Regeln auf eine vorher festgelegte Anzahl zu reduzieren. Möglich wäre zum einen die inhaltliche Zusammenfassung der Regeln, welche die Anzahl der vorhandenen Literale in den verschiedenen Regeln berücksichtigt. Zum anderen könnten die Vorarbeiten aus Kapitel 4.5 verwendet werden, um die Regeln nach ihrer Signifikanz zu bündeln.

Eine inhaltliche Zusammenfassung der Regeln setzt voraus, dass einzelne oder mehrere Literale in verschiedenen Implikationen vorhanden sind und daher zu einer Regel um Literal x_j oder um die Literale $(x_i \wedge x_j)$ zusammengefasst werden könnten. Zum Auffinden solcher Literale könnte die Häufigkeit ihres Auftretens untersucht werden. Allerdings wird die Signifikanz der Regeln bei einem solchen Vorgehen vernachlässigt. Weiterhin müsste geklärt werden, wieviele Regeln eine zusammengefasste Teilmenge enthalten soll. Dies ist im Hinblick auf eine gleichmäßige Verteilung der Datensätze auf die Teilmengen von Bedeutung.
Hinsichtlich unserer Zielformulierung wirft die inhaltliche Zusammenfassung der Regeln einige Probleme auf. Die Voraussetzung, dass Literale in verschiedenen Regeln vorhanden sind, ist nicht immer gegeben, wie Beispiel 4.6 verdeutlichen soll.

Beispiel 4.6. *Gegeben seien die Regeln r_i, $i = \{1,2,3,4\}$ mit insgesamt neun Literalen, die zusammengefasst werden sollen.*

	x_1	x_2	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
r_1	1	1								
r_2			1	1			1			
r_3					1	1		1		
r_4									1	1

$\Rightarrow$ *Da die Literale nicht gleichzeitig in mehreren Regeln vorhanden sind, kann keine inhaltliche Zusammenfassung durchgeführt werden.*

Daneben können verschiedene Literale genauso häufig in unterschiedlichen Regeln auftauchen. Ein reiner Vergleich der Häufigkeit des Auftretens liefert hier ebenfalls kein eindeutiges Ergebnis. Zudem besteht das Problem, dass Literale in Regeln mit deutlich unterschiedlicher Signifikanz auftreten können und damit die Sinnhaftigkeit

einer solchen Zusammenfassung in Frage gestellt wird. Die Zusammenfassung unterschiedlicher Signifikanzen könnte zudem die Klassifikationsergebnisse verschlechtern. Die hier angeführten Argumente legen den Schluss nahe, dass die inhaltliche Zusammenfassung nicht stets zur gewünschten Regelreduktion führt und daher für unsere Zwecke als eher unbrauchbar einzuschätzen ist.

Bei der Signifikanzanalyse in Unterabschnitt 4.5.3 wurden Vorarbeiten durchgeführt, die auch für eine Zusammenfassung verwendet werden könnten. Allerdings ist zu klären, ob mit Hilfe des dort ermittelten χ^2–Wertes direkt auf die Stärke des Zusammenhangs geschlossen werden kann. Dies ist generell nicht möglich, da der χ^2–Wert von der Stichprobengröße abhängig ist und sich daher mit steigender Stichprobengröße auch ein höherer χ^2–Wert ergibt. Ein hoher χ^2–Wert kann aus schwachen Zusammenhängen resultieren, wenn die Stichprobengröße nur ausreichend hoch gewählt wird. Daher werden im folgenden Abschnitt verschiedene Zusammenhangsmaße kurz vorstellen.

4.6.3 Zusammenhangsmaße

Die Statistik gibt verschiedene Zusammenhangsmaße für Nominalskalen an, die auf dem χ^2–Konzept beruhen [Sac02, BB96]. Diese versuchen den Wert so zu normieren, dass er von der Stichprobengröße unabhängig ist. Der einzige Unterschied der verschiedenen Maße liegt in der Art der Normierung. Dabei bezeichnet N stets den gesamten Stichprobenumfang.

- Kontingenzkoeffizient C (Kontingenzkoeffizient von Pearson)

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

  Der Koeffizient C ist ein Maß für die Straffheit des Zusammenhangs.

- Phi

$$\Phi = \sqrt{\frac{\chi^2}{N}}$$

  Die Größe Φ ist eine weitere Maßzahl für die Stärke des Zusammenhangs.

- Cramers V

$$V = \sqrt{\frac{\chi^2}{N \cdot (min(r,c) - 1)}}$$

  Beim Zusammenhangsmaß V bezeichnet r die Anzahl der Spalten und c die Anzahl der Zeilen. Für eine Vierfeldertafel entspricht V dem Maß Φ.

Die Koeffizienten werden alle mit dem Stichprobenumfang normiert. Da für unsere Modellierung zur Ermittlung der Signifikanz stets eine 2×2 Kontingenztabelle und derselbe Stichprobenumfang verwendet wird, kann zur Messung der Stärke des Zusammenhangs direkt der χ^2–Wert verwendet werden. Durch Anwendung einer der hier vorgestellten Maßzahlen würde sich lediglich die Normierung ändern.

4.6.4 Zusammenfassung mit Hilfe von Lageparametern

Gegeben seien die Regeln beider Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ für die Mengen G^+ und G^- mit zugehörigen nicht normierten χ^2–Werten, die anhand dieses Wertes sortiert und in eine Rangfolge gebracht werden können.
Ziel ist es, eine Zusammenfassung der Regeln zu finden, die stets durchführbar und einfach zu implementieren ist und zudem die Stärke der einzelnen Regeln berücksichtigt. Im Hinblick auf die Anwendung des Modells im Rahmen von IRB–Ansätzen, soll die Zusammenfassung weiterhin eine vorher festgelegte Anzahl von Teilmengen erzeugen, die anschließend mit dem vorhandenen Datenmaterial empirisch bewertet werden können. Dies entspricht der Bildung unterschiedlicher Bonitätsklassen, auf die im zweiten Teil dieser Arbeit ausführlich eingegangen wird.

Bei der Vielzahl der vorhandenen Regeln ist eine solche Umsetzung schwer durchführbar, daher wird eine stark vereinfachte Vorgehensweise vorgeschlagen. Die Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ liegen in Form einer Rangfolge ihres χ^2–Wertes vor. Gemäß der Zielformulierung soll eine vorab festgelegte Anzahl an Teilmengen entstehen. Dies kann auf einfache Weise durch die Verwendung von Lageparametern, welche die Signifikanz der Regeln berücksichtigen, erreicht werden. Einen geeigneten Lageparameter stellt der Median dar, da durch ihn stets eine Reduzierung beider Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ erreicht wird. Zusätzlich bietet die Zusammenfassung mit Hilfe des Medians den Vorteil, dass die entstandenen Teilmengen stets dieselbe Anzahl an Regeln enthalten. Damit wird der gewünschten gleichmäßigen Verteilung der Datensätze auf alle Teilmengen Rechnung getragen.

Bei Vorlage einer geordneten Messreihe $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ mit $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ ergibt sich für den Median x_{med} bei einer ungeraden Beobachtungsanzahl $n : x_{med} = x_{(\frac{n+1}{2})}$, bei einer geraden Beobachtungsanzahl $n : x_{med} = x_{(\frac{n}{2})}$. Durch die Verwendung des Medians wird also stets eine Halbierung der Mengen und daraus resultierend eine feste Anzahl zu bewertender Teilmengen erzeugt. Dabei bezeichnet

$\mathcal{L}_{1\bullet}$ = Regeln, die aus dem Kontext $\mathbb{K}^+$ ermittelt wurden,
$\mathcal{L}_{2\bullet}$ = Regeln, die aus dem Kontext $\mathbb{K}^-$ ermittelt wurden.

Die Zusammenfassung mit Hilfe des Medians liefert in Anlehnung an Tabelle 4.4 fol-

gende Regeln in disjunktiver Normalform für die Menge G^+

$$\begin{aligned}\mathcal{L}_{11} &= r_{(1)} \vee r_{(2)} \vee \ldots \vee r_{(\frac{n}{2})} \quad \text{und} \\ \mathcal{L}_{12} &= r_{(\frac{n}{2}+1)} \vee r_{(\frac{n}{2}+2)} \vee \ldots \vee r_{(n)}.\end{aligned}$$

Für die Signifikanz der beiden zusammengefassten Regeln gilt $\mathcal{L}_{11} > \mathcal{L}_{12}$. Analoges gilt für die Darstellung der Regeln der Menge G^-

$$\begin{aligned}\mathcal{L}_{21} &= \hat{r}_{(1)} \vee \hat{r}_{(2)} \vee \ldots \vee \hat{r}_{(\frac{m}{2})} \quad \text{und} \\ \mathcal{L}_{22} &= \hat{r}_{(\frac{m}{2}+1)} \vee \hat{r}_{(\frac{m}{2}+2)} \vee \ldots \vee \hat{r}_{(m)}.\end{aligned}$$

Für die Signifikanz gilt ebenfalls $\mathcal{L}_{21} > \mathcal{L}_{22}$. Dabei gilt eine zusammengefasste Regel $\mathcal{L}_{ij}$, $i, j = \{1,2\}$ als erfüllt, wenn eine ihrer Klauseln erfüllt ist. Die kombinatorische Teilmengenbildung erlaubt die Bildung von $|\mathcal{P}(k)| = |\mathcal{P}(4)| = 2^4 = 16$ Teilmengen. Möglich ist z. B. die Regel

$$\begin{aligned}(\mathcal{L}_{11}\mathcal{L}_{12}) &= \mathcal{L}_{11} \wedge \mathcal{L}_{12} \\ &= (r_{(1)} \vee r_{(2)} \vee \ldots \vee r_{(\frac{n}{2})}) \wedge (r_{(\frac{n}{2}+1)} \vee r_{(\frac{n}{2}+2)} \vee \ldots \vee r_{(n)})\end{aligned}$$

in konjunktiver Normalform, die erfüllt ist, wenn $\mathcal{L}_{11}$ und $\mathcal{L}_{12}$ erfüllt sind. Die Zusammenfassung und Kombination der Regeln liefert eine natürliche Zerlegung des Bestandes in unterschiedliche Klassen. Dies ist vor allem im Hinblick auf die Anwendung des Modells im Rahmen von IRB–Ansätzen hilfreich (siehe Abschnitt 5.8). Zusätzlich kann bei Anwendung des Modells im Rahmen der Klassifikation gezielt jede vom Datensatz erfüllte Regel mit ihrem zugehörigen χ^2–Wert betrachtet werden.

Trotz der erwähnten Nachteile einer differenzierten inhaltlichen Zusammenfassung der Regeln wurde die Klassifikation am Beispiel der Kreditausfälle sowohl mit einer inhaltlichen, als auch mit der Medianzusammenfassung durchgeführt. Dabei zeigte sich, dass eine differenzierte Zusammenfassung der Regeln keine wesentliche Verbesserung in den Klassifikationsergebnissen liefert. Daher wurde in den Anwendungsbeispielen im zweiten Teil der Arbeit stets das standardisierte Verfahren mit Hilfe des Medians gewählt. Allerdings müssen die Mengen $\mathcal{L}_{ij}$, $i, j = \{1,2\}$ noch mit Hilfe empirischer Trainingsdaten bewertet werden. Dieses Problem wird im folgenden Abschnitt durch die Verwendung bedingter Wahrscheinlichkeiten gelöst.

4.7 Bewertung der Regeln

Aus den umfangreichen Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ sind mit Hilfe des Medians vier Teilmengen $\mathcal{L}_{11}, \mathcal{L}_{12}, \mathcal{L}_{21}$ und $\mathcal{L}_{22}$ entstanden, die jeweils einen Teil der Regeln umfassen. Abbildung 4.1 zeigt die möglichen Regelkombinationen, die aus der Medianzusammenfassung entstanden sind. Die leere Menge bezeichnet dabei den Fall, dass der zu

$$\emptyset$$

$$(\mathcal{L}_{11}) \quad (\mathcal{L}_{12}) \quad (\mathcal{L}_{21}) \quad (\mathcal{L}_{22})$$

$$(\mathcal{L}_{11}\mathcal{L}_{12}) \quad (\mathcal{L}_{11}\mathcal{L}_{21}) \quad (\mathcal{L}_{11}\mathcal{L}_{22}) \quad (\mathcal{L}_{12}\mathcal{L}_{21}) \quad (\mathcal{L}_{12}\mathcal{L}_{22}) \quad (\mathcal{L}_{21}\mathcal{L}_{22})$$

$$(\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}) \quad (\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{22}) \quad (\mathcal{L}_{11}\mathcal{L}_{21}\mathcal{L}_{22}) \quad (\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22})$$

$$(\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22})$$

Abbildung 4.1: Mögliche Regelkombinationen nach der Medianzusammenfassung der Ursprungsmengen

untersuchende Datensatz keine Regel erfüllt. Es muss allerdings noch ein Bewertungsverfahren für die Teilmengen gefunden werden. Ein Datensatz wird nun genau der maximalen Menge der erfüllten Regeln zugeordnet. Erfüllt Datensatz k_i beispielsweise die Regeln $\mathcal{L}_{11}$ und $\mathcal{L}_{21}$, so wird er ausschließlich der Vereinigung der beiden Mengen zugeordnet. Dadurch ist eine eindeutige Zuordnung der Datensätze zu den Teilmengen gewährleistet. Diesen Sachverhalt können wir mit Hilfe eines einfachen Entscheidungsbaumes darstellen. Abbildung 4.2 zeigt den zugehörigen Entscheidungsbaum für unsere Fragestellung am Beispiel der Teilmenge $(\mathcal{L}_{11}\mathcal{L}_{21})$.

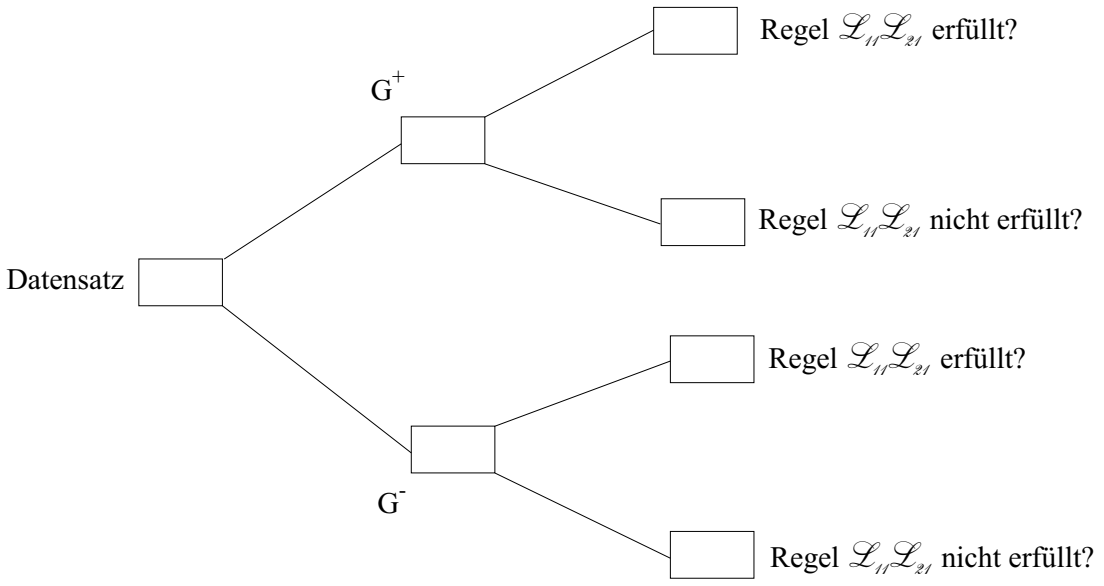


Abbildung 4.2: Entscheidungsbaum zur Ermittlung der bedingten Wahrscheinlichkeiten

Für jeden Datensatz der die Regeln $\mathcal{L}_{11}\mathcal{L}_{21}$ erfüllt, existieren zwei mögliche Pfade

im Entscheidungsbaum: Entweder er stammt aus der Menge G^+ oder er stammt aus der Menge G^-. Analoges gilt für jeden Datensatz k_i und jede der 16 Teilmengen aus Tabelle 4.1.

Im Rahmen der Wahrscheinlichkeitstheorie lässt sich unsere Fragestellung auch folgendermaßen formulieren. Für i vorliegende Datensätze soll überprüft werden, ob sie bestimmte Regeln erfüllen oder nicht. Dabei bezeichnen wir Ω als Ergebnisraum der alle Datensätze umfasst, ω nennen wir Elementarereignis, mit

$$\begin{aligned} \omega_i &= 1 : \text{Datensatz } k_i \text{ erfüllt die Regel,} \\ \omega_i &= 0 : \text{Datensatz } k_i \text{ erfüllt die Regel nicht.} \end{aligned}$$

Ein Ereignis A bezeichnet in unserem Fall das Vorliegen einer bestimmten Teilmenge aus $\mathcal{P}(\mathcal{L}_{ij})$. Der Ergebnisraum Ω kann damit folgendermaßen beschrieben werden:

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots, \omega_n) : \omega_i = 0 \text{ oder } 1, \quad i = 1, \ldots, n\}.$$

Allerdings besitzen wir noch weitere Informationen über unser Zufallsexperiment. Wir wissen, dass der Ergebnisraum Ω durch die disjunkten Mengen G^+ und G^- vollständig zerlegt wird. Daher gilt:

$$\begin{aligned} G^+ \cap G^- &= \emptyset \quad \text{und} \\ G^+ \cup G^- &= \Omega. \end{aligned}$$

Daneben besitzen wir Kenntnisse über die empirischen Größen G^+ und G^-, für die gilt:

$$\begin{aligned} |G^+| &= g \\ |G^-| &= k. \end{aligned}$$

Damit gelten die Voraussetzungen aus Satz 2.9 und wir können mit Hilfe der Formel von Bayes Bewertungen für jede Teilmenge ermitteln. So erhalten wir für jede Teilmenge eine bedingte Wahrscheinlichkeit dafür, dass der Datensatz aus der Menge G^+ oder G^- stammt, wenn Ereignis A vorliegt. Formal werden diese Wahrscheinlichkeiten mit $P(G^+|A)$ bzw. $P(G^-|A)$ bezeichnet und folgendermaßen ermittelt:

$$\begin{aligned} P(G^+|A) &= \frac{P(G^+) \cdot P(A|G^+)}{P(G^+) \cdot P(A|G^+) + P(G^-) \cdot P(A|G^-)} \\ P(G^-|A) &= \frac{P(G^-) \cdot P(A|G^-)}{P(G^+) \cdot P(A|G^+) + P(G^-) \cdot P(A|G^-)} \end{aligned}$$

Dabei bezeichnen wir $P(G^+)$ und $P(G^-)$ als a priori Wahrscheinlichkeiten, die unser Vorwissen über die Verteilung der Mengen G^+ und G^- widerspiegeln. Die bedingten

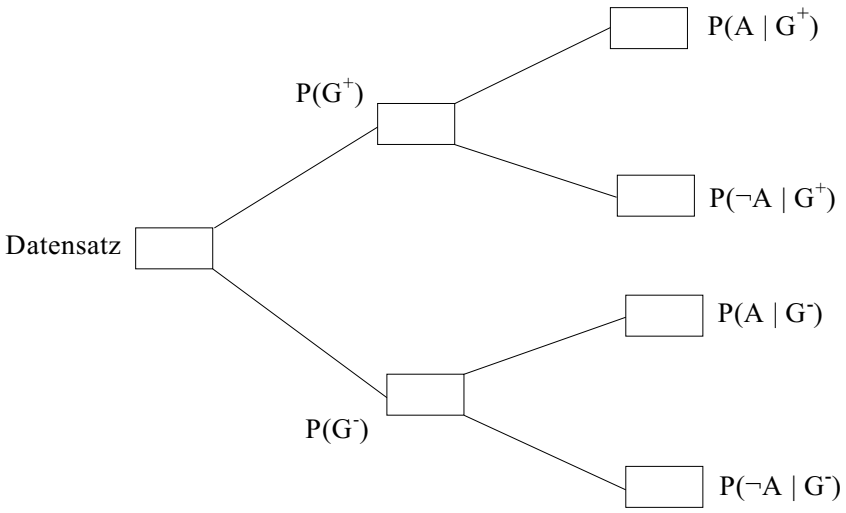


Abbildung 4.3: Präzisierter Entscheidungsbaum zur Ermittlung der bedingten Wahrscheinlichkeiten

Wahrscheinlichkeiten $P(G^+|A)$ und $P(G^-|A)$ nennt man a posteriori Wahrscheinlichkeiten. Sie geben an, wie hoch die Wahrscheinlichkeit ist, dass der Datensatz aus der Menge G^+ bzw. G^- stammt, wenn Ereignis A im Datensatz k_i eintritt. Damit kann unser Entscheidungsbaum aus Abbildung 4.2 präzisiert werden. Wir sind nun in der Lage für alle 16 Teilmengen bedingte Wahrscheinlichkeiten zu ermitteln. Diese Wahrscheinlichkeiten liefern uns eine Bewertung, mit deren Hilfe wir die Datensätze klassifizieren können. Allerdings müssen die Bewertungen und die daraus resultierenden Klassifikationsergebnisse geeignet interpretiert und ausgewertet werden. Dazu verwenden wir die ROC–Analyse, auf die wir im folgenden Abschnitt näher eingehen wollen.

4.8 Analyse der Ergebnisse

4.8.1 ROC–Graphen

Zur Analyse der Ergebnisse werden die in 2.5 vorgestellten Verfahren verwendet. Mit Hilfe von Sensitivität und Spezifität sind wir in der Lage einen ROC–Graphen zu erstellen. Sensitivität und Spezifität werden folgendermaßen ermittelt

$$\text{Sensitivität} = \frac{TP}{TP+FN} \qquad \text{Spezifität} = \frac{TN}{TN+FP}.$$

Generell stellt sich bei allen Klassifikationsmodellen die Frage, ab welchem Schwellwert ein Datensatz positiv bzw. negativ zu bewerten ist. Eine Ausnahme bilden die

diskreten Klassifikationsverfahren wie die Entscheidungsbäume. Sie liefern ein eindeutiges Ergebnis, welches einem einzelnen Punkt im ROC–Graphen entspricht. In [Faw03] wird ein Verfahren vorgestellt, das die Erstellung einer ROC–Kurve aus Entscheidungsbäumen möglich macht und damit auch die Ermittlung eines charakteristischen AUC–Wertes.

Die Angabe von Wahrscheinlichkeiten stellt einen Grad der Zugehörigkeit eines Datensatzes zu einer bestimmten Klasse dar. Der Grad der Zugehörigkeit kann zusammen mit einem Schwellwert $t \in \mathbb{R}$ verwendet werden, um ein diskretes Klassifikationsergebnis zu erhalten. Ist der Grad der Zugehörigkeit größer t, so wird der Datensatz der Klasse G^+ zugeordnet, ist er kleiner t, wird er der Menge G^- zugewiesen. Bei Angabe von Wahrscheinlichkeiten liegen die Schwellwerte im Intervall $[0,1]$. Die Addition der Stimmen pro Datensatz liefert z. B. im Modell von Truemper Schwellwerte zwischen $[-40,40]$. Generell können die Schwellwerte im Intervall $[-\infty,\infty]$ liegen und ermöglichen verschiedene Zuordnungen zu den jeweiligen Klassen. Mit Hilfe der Schwellwerte können unterschiedliche Kontingenztabellen erstellt, und daraus resultierend verschiedene Werte für Sensitivität und Spezifität ermittelt werden.

In [Faw03] wird ein einfacher Algorithmus mit Laufzeit $O(n^2)$ zur Erstellung einer ROC–Kurve beschrieben. Er prüft für alle positiven Datensätze, ob die Wahrscheinlichkeiten $f(i)$ über dem Schwellwert t liegen (Algorithmus 1, Zeile 5) und ordnet ihn entweder der Variable TP oder FP zu. Dabei werden alle n Datensätze n mal durchlaufen.
Der Algorithmus kann effizienter gestaltet werden, wenn die Monotonieeigenschaft der Ergebnisse ausgenutzt wird. Dazu werden vorab die vorliegenden rationalen Klassifikationsergebnisse absteigend sortiert. Wird ein Datensatz für einen Schwellwert positiv eingestuft, so ist dies für niedrigere Schwellwerte ebenfalls der Fall. Die Liste wird nach unten abgearbeitet und die Anzahl der TP stets erhöht. Zur Sortierung benötigt der Algorithmus die Laufzeit $O(n \cdot \log n)$ sowie $O(n)$ um die Liste nach unten abzuarbeiten. Daraus resultiert eine Gesamtkomplexität von $O(n \cdot \log n)$ [Faw03].

4.8.2 AUC–Werte

Zur Analyse der Ergebnisse wollen wir zusätzlich den AUC–Wert angeben, der mit Hilfe einiger Änderungen des oben vorgestellten Algorithmus errechnet werden kann. Algorithmus 2 zeigt ein effizientes Verfahren zur Ermittlung des AUC–Wertes. Dabei werden sukzessive die Flächen der einzelnen Trapeze hinzugefügt. Voraussetzung des Algorithmus ist ebenfalls die absteigende Sortierung der Datensätze nach ihren f–Werten.
In [Faw03] ist das Vorgehen zum Vergleich zweier Klassifikationsmodelle anhand ihrer ROC–Graphen ausführlich dargestellt. Um einen umfassenden Vergleich zweier Klassifikationsmodelle zu erlangen, sollten ROC–Kurven aus Datensätzen verschiedener

Algorithmus 1 : Einfache Methode zur Generierung einer ROC–Kurve

Input : L = Menge aller Testdatensätze, $f(i)$ = Wahrscheinlichkeit, dass die Instanz i positiv ist, *min* und *max* = der kleinste bzw. größte Wert der von f ausgegeben wird, *increment* = kleinste Differenz zwischen zwei beliebigen Werten von f, P = Anzahl der positiven Trainingsbeispiele, N = Anzahl der negativen Trainingsbeispiele

Output : ROC–Kurve

```
1  for t = min to max by increment do
2      FP ← 0
3      TP ← 0
4      for i ∈ L do
5          if f(i) > t then
6              if i ist ein positiver Datensatz then
7                  TP ← TP +1
8              else
9                  FP ← FP +1
10             end
11         end
12     end
13     Punkt (FP/N, TP/P) zur ROC–Kurve hinzufügen
14 end
```

Testmengen erstellt werden. Daraus kann z. B. mit Hilfe des „vertical averaging" oder des „threshold averaging" eine gemittelte ROC–Kurve erstellt werden. Eine ausführliche Vorgehensweise beider Methoden ist in [MP04] dargestellt.
Wir werden im zweiten Teil der Arbeit zum Vergleich der Klassifikationsmodelle die ROC–Kurven und die AUC–Werte aus verschiedenen Testmengen verwenden. Die Ergebnisse des verbandstheoretischen Implikationenmodell sollen mit den Ergebnissen der neuronalen Netze, der Entscheidungsbäume und des Modells von Truemper verglichen werden. Dabei werden wir auf eine Mittelung der ROC–Kurven verzichten, da die Schwankungen innerhalb der Modelle äußerst gering waren und wir zudem am generellen Klassifikationsniveau der Modelle interessiert sind. Weiterhin liegt unser Augenmerk auf dem qualitativen Vergleich der Modelle. Zur spezifischen Analyse wurden die Ergebnisse der Modelle mit der höchsten Klassifikationsgüte gewählt.

Im folgenden Abschnitt wird das vorgestellte verbandstheoretische Implikationenmodell dazu verwendet, ausgefallene Bauspardarlehen zu klassifizieren und Kreditausfallwahrscheinlichkeiten zu ermitteln. Darauf aufbauend wird eine Verwendung des Modells im Rahmen von IRB–Ansätzen vorgestellt. Die erhaltenen Klassifikationsergebnisse werden mit den Ergebnissen der in Kapitel 3 vorgestellten Modelle

verglichen. Zudem wurde bei allen Modellen die Generalisierungsfähigkeit überprüft, indem sie ohne weitere Anpassungen auf die Originaldaten einer weiteren Bausparkasse angewendet wurden. In einem letzten Schritt werden weitere bauspartechnische Fragestellungen anhand realer Kollektivdaten mit Hilfe des verbandstheoretischen Implikationenmodells untersucht.

Algorithmus 2 : Ermittlung des Flächeninhalts unter einer ROC–Kurve

Input : L = Menge aller Testdatensätze, $f(i)$ = Wahrscheinlichkeit, dass die Instanz i positiv ist, P = Anzahl der positiven Trainingsbeispiele, N = Anzahl der negativen Trainingsbeispiele

Output : AUC–Wert, Flächeninhalt unter der ROC–Kurve

1 $L_{sorted} \leftarrow L$ absteigend nach f Werten sortiert
2 $FP \leftarrow TP \leftarrow 0$
3 $FP_{prev} \leftarrow TP_{prev} \leftarrow 0$
4 $A \leftarrow 0$
5 $f_{prev} \leftarrow -\infty$
6 $i \leftarrow 1$
7 **while** $i \leq |L_{sorted}|$ **do**
8 **if** $f(i) \neq f_{prev}$ **then**
9 $A \leftarrow A +$ TRAPEZOID_FLÄCHE$(\frac{FP}{N}, \frac{FP_{prev}}{N}, \frac{TP}{P}, \frac{TP_{prev}}{P})$
10 $f_{prev} \leftarrow f(i)$
11 $FP_{prev} \leftarrow FP$
12 $TP_{prev} \leftarrow TP$
13 **end**
14 **if** *i ist ein positiver Datensatz* **then**
15 $TP \leftarrow TP + 1$
16 **else**
17 $FP \leftarrow FP + 1$
18 **end**
19 **end**
20 $A \leftarrow A +$ TRAPEZOID_FLÄCHE$(1, \frac{FP_{prev}}{N}, 1, \frac{TP_{prev}}{P})$
21 `Funktion(`*TRAPEZOID_FLÄCHE*(X_1, X_2, Y_1, Y_2)
22 *Breite* $\leftarrow |X_1 - X_2|$
23 *Höhe* $\leftarrow (Y_1 + Y_2)/2$
24 ***return*** [Breite x Höhe])

Teil II

Anwendung

Kapitel 5

Ermittlung von Kreditausfallwahrscheinlichkeiten mit dem verbandstheoretischen Implikationenmodell

5.1 Bisherige Untersuchungen und Eigenarten von Privatkundenkrediten

Im Bereich der Kreditwürdigkeitsprüfung von Privatkunden und Unternehmen existieren zahlreiche Untersuchungen mit unterschiedlichen Modellierungsansätzen. Beispiele finden sich unter anderem in [MR00, KS00, Füs95, Sch03, OU02, SR94]. Der Großteil der Untersuchungen beschäftigt sich mit der Bonitätsanalyse von Unternehmen. Den Untersuchungen liegen dann meist Größen wie Bilanzkennzahlen, Umsatzzahlen, Jahresabschlüsse usw. zugrunde. Untersuchungen im Privatkundenbereich sind meist auf persönliche oder sozio–demographische Merkmale angewiesen [SR94]. Dies hat mehrere Nachteile:

- Fehlende Zeitachse
- Geringere Variabilität
- Nachprüfbarkeit
- Fehlende Angaben

Bei Privatkundenkrediten stehen Merkmale, wie z. B. Einkommen, Beruf, Darlehenshöhe und Alter im Vordergrund, die bei Vergabe des Darlehens ausgewertet werden. Die Untersuchung von reinen Bauspardarlehen bietet im Vergleich dazu den Vorteil, dass zusätzliche Merkmale aus dem bisherigen Verlauf des Bausparvertrages in die Analysen miteinbezogen werden können (z. B. Spardauer, WoP–Anteil am Guthaben

oder Tarif). Es handelt sich nicht mehr um eine rein stichtagsbezogene Abfrage von Merkmalen, sondern vielmehr werden Vergangenheitswerte in die Bonitätsprüfung miteinbezogen. Im Verlauf dieses Kapitels soll untersucht werden, inwieweit sich dadurch die Prognosequalität der Aussagen über einen Ausfall bzw. Nichtausfall eines Bauspardarlehens verbessern lässt. Zusätzlich bietet die Verwendung von Bauspardarlehen in unserem Fall den Vorteil, dass fehlende Werte bei der Vielzahl der vorliegenden realen Kollektivdaten eine untergeordnete Rolle spielen. Einige der oben angeführten Nachteile bei der Bonitätsanalyse von Privatkundenkrediten können durch die Verwendung von Bauspardarlehen beseitigt werden.

Zur Kreditwürdigkeitsuntersuchung werden daher wirtschaftliche, persönliche sowie vertragsimmanente Merkmale herangezogen. In den folgenden Abschnitten wird die Anwendung des verbandstheoretischen Implikationenmodells zur Klassifikation von ausgefallenen bzw. endgetilgten Bauspardarlehen vorgestellt. In einem weiteren Schritt soll überprüft werden, inwieweit die ermittelten Ausfallwahrscheinlichkeiten im Rahmen von IRB–Ansätzen verwendet werden können. Vorab wird der Begriff der Kreditausfallwahrscheinlichkeit näher definiert.

5.2 Definition Kreditausfallwahrscheinlichkeit

Um die schuldnerspezifischen Kreditrisiken zu ermitteln werden Ausfallwahrscheinlichkeiten bestimmt. Unter Kreditausfallwahrscheinlichkeit versteht man die Wahrscheinlichkeit, dass ein Darlehensnehmer nicht in der Lage ist, seinen Zahlungsverpflichtungen bestehend aus Zins und Tilgung nachzukommen und somit ausfällt:

$$\pi(x) \quad = \quad P(Y_j = 1 | X_j = x), \quad x \in \mathcal{X} \tag{5.1}$$

$X_j \in \mathbb{R}^d$ steht dabei für die bei Kreditvergabe vorhandenen Merkmale des j–ten Darlehensnehmer mit dem Wertebereich $\mathcal{X}$. Allerdings müssen die vorhandenen Merkmale für eine Verwendung im verbandstheoretischen Implikationenmodell binär kodiert werden. Dies wird in Unterabschnitt 5.6.2 erläutert. Für die Indikatorvariable des Kredits $Y_j \in \{0,1\}$ gilt:

$$Y_j \quad = \quad \begin{cases} 1; & \text{Schuldner } j \text{ fällt aus} \\ 0; & \text{Schuldner } j \text{ fällt nicht aus} \end{cases} \tag{5.2}$$

Die Ausfallwahrscheinlichkeit $\pi(x)$ nimmt Werte im Intervall $[0,1]$ an. Im Rahmen dieser Arbeit werden ausschließlich Kreditausfälle bzw. Nichtausfälle und keine Bonitätsverschlechterungen im Zeitverlauf betrachtet.

5.3 Untersuchungsaufbau

Für die Untersuchung wurden reale Kollektivdaten einer Bausparkasse der Jahre 2000–2004 verwendet, die in Form von Jahresendbeständen vorliegen. Flussgrößen, wie der monatliche Spar- oder Tilgungseingang, sind dabei aufsummiert. Der Datenumfang umfasst ca. drei Millionen Bausparkonten. Allerdings sind für unsere Zwecke nur Bausparkonten von Interesse, die sich aktuell in der Darlehensphase befinden bzw. diese bereits beendet haben. In den Originaldaten der Ausgangsbausparkasse befinden sich ca. 350.000 Bausparverträge, die diesen Bedingungen entsprechen. Davon ist jedoch lediglich ein Teil für unsere Untersuchungen relevant, wie im Folgenden erläutert wird. Bei den Bausparkassen ist grundsätzlich zwischen kollektivem und außerkollektivem Geschäft zu unterscheiden. Zum kollektiven Geschäft zählen die klassischen Bauspardarlehen, wohingegen Vor- und Zwischenfinanzierungsverträge in den außerkollektiven Geschäftsbereich fallen. Bei klassischen Bauspardarlehen entsteht ein Zahlungsverzug in der Darlehensphase. Im Gegensatz dazu entsteht ein Zahlungsverzug bei VK/ZK–Verträgen bereits in der Sparphase. Beim Vorfinanzierungsvertrag wird von der Bausparkasse ein Vorausdarlehen gewährt. Im Gegenzug wird vom Darlehensempfänger ein Bausparvertrag über die volle Summe des Vorausdarlehens abgeschlossen. Der Bausparer ist verpflichtet, den Bausparvertrag mit einer vertraglich festgelegten Rate zu besparen. Die Sparleistung bei vorfinanzierten Verträgen entspricht daher der Tilgungsleistung der klassischen Bauspardarlehen. Da der Zahlungsverzug bereits in der Sparphase entsteht, werden VK/ZK–Verträge nicht in die Untersuchung miteinbezogen, es werden ausschließlich reine Bauspardarlehen betrachtet.
Eine weitere Einschränkung besteht in der Auswahl der Darlehenskonten. Im Rahmen unserer Untersuchung sind bereits ausgefallene, sowie endgetilgte Darlehen von besonderem Interesse. Bauspardarlehen, die sich aktuell noch in der Darlehensphase befinden, werden nicht für die Untersuchung verwendet, da sie im Verlauf ihrer weiteren Darlehensphase noch in Zahlungsverzug geraten können. Diese Problematik bereitet bei Kreditwürdigkeitsprüfungen mit geringem Datenmaterial häufig Schwierigkeiten. In den vorliegenden Originaldaten finden sich jedoch genügend Datensätze, die unseren Voraussetzungen genügen. Das vorhandene Datenmaterial wurde in eine Trainings- und Testmenge im Verhältnis 1:1 aufgeteilt, da eine Abweichung von diesem Verhältnis stets zu schlechteren Klassifikationsergebnissen führte (siehe Unterabschnitte 6.1.3 und 6.2.2). Zur Ermittlung des Regelwerks und der Ausfallwahrscheinlichkeiten wurden aus den Trainingsdaten verschiedene Stichproben gezogen, zur Validierung der Ergebnisse wurden die Testdaten verwendet.

5.4 Darlehensauswahl

Die Klassifikation der Bauspardarlehen erfolgt stichtagsbezogen zum Jahresende. Für die Untersuchung wurden ausschließlich ältere Tarifgenerationen betrachtet, da die

Anzahl der endgetilgten Verträge in diesen Tarifen ausreichend hoch ist. Bausparkonten neuerer Tarifgenerationen befinden derzeit überwiegend in der Sparphase. Die Auswahlkriterien werden in den folgenden Abschnitten näher erläutert.

5.4.1 Ausgefallene Darlehen

Bauspardarlehen, die im Untersuchungszeitraum jemals einen Zahlungsverzug von mehr als 90 Tagen aufwiesen, werden als ausgefallen eingestuft. Dieses Kriterium wird auch vom Baseler Ausschuss für Bankenaufsicht in der Rahmenvereinbarung [Bas04][1] vorgeschlagen. Ein Zahlungsrückstand wird auf Basis der von der Bausparkasse vertraglich festgelegten Mindesttilgung ermittelt, die abhängig vom jeweiligen Tarif ist. Tabelle 5.1 zeigt die Tilgungsraten (tr) der zugrunde liegenden Tarife der Ausgangsbausparkasse. Dabei beziehen sich die Tilgungsraten stets auf die volle Bausparsumme (bs). Zur Ermittlung der Regeltilgungszeit wird außerdem der Anspargrad des Tarifes benötigt. Für einen Darlehensausfall muss bei Vorlage von Jahresdaten für das Daten-

Tarifname	Anspargrad	Tilgungsrate p.a. (tr)
Tarif 1	40 %	7.2 %
Tarif 2	50 %	9.6 %
Tarif 3	40 %	7.2 %
Tarif 4	50 %	6.0 %

Tabelle 5.1: Tarifkonditionen der Alttarife der Ausgangsbausparkasse

feld „Tilgungsbeitrag in Euro" (tb) folgende Bedingung gelten:

$$tb < tr \cdot 0.75 \cdot bs$$

Für Bauspardarlehen, die dieser Bedingung genügen, wird die Indikatorvariable $Y_i = 1$ (Schuldner i fällt aus) gesetzt. Die Auswahlbedingungen lieferten 2.730 ausgefallene Darlehenskonten, mit einem gesamten Darlehensvolumen von ca. zehn Millionen Euro im Bestand der Ausgangsbausparkasse.

5.4.2 Endgetilgte Darlehen

Ein Darlehen wird als nicht ausgefallen bewertet, wenn das Bauspardarlehen innerhalb der Regeltilgungszeit endgetilgt wurde. Wir betrachten daher Konten, die im Untersuchungszeitraum ihr Darlehen innerhalb der Regelzeit endgetilgt haben und von der Bausparkasse bereits abgewickelt sind. Die Bausparkonten müssen daher im Datenfeld „Auflösungsgrund" den Eintrag „Darlehensende" aufweisen.
Aufgrund des beschränkten Zeithorizonts und der Vorlage von Jahresdaten ist bei endgetilgten Bauspardarlehen nicht die komplette Darlehensphase ersichtlich. Theoretisch

[1] Vgl. Abschnitt 452 in [Bas04].

können endgetilgte Darlehen unterjährig oder vor der „sichtbaren" Darlehensphase in Zahlungsverzug geraten sein. Allerdings hätten dann Nachzahlungen der Tilgungsraten stattfinden müssen. Zur Verringerung dieses Risikos wird ein zusätzlicher Sicherheitspuffer verwendet, indem die ermittelten Regellaufzeiten stets nach „unten abgerundet" werden. Die ausgewählten Darlehen tilgen daher immer schneller als von der Bausparkasse gefordert. Da die Regeltilgungszeiten der Darlehen abhängig vom Tarif sind, werden diese tarifspezifisch mit Hilfe der Mindesttilgungsraten aus Tabelle 5.1 ermittelt.

Beispiel 5.1. *Ermittlung der Regellaufzeit für den Tarif 1 unter der Annahme, dass der Bausparer die maximal mögliche Darlehenssumme in Anspruch nimmt.*

Anspargrad Tarif 1: 40 %
Tilgungsrate (tr) p.a.: 7.2 %
Darlehenslaufzeit bei Regeltilgung des Darlehens (in Jahren): x

$$\begin{aligned} 0 &= bs \cdot 0.6 - x \cdot (tr \cdot bs) \\ 0 &= bs \cdot 0.6 - x \cdot (0.072 \cdot bs) \\ &\approx 8.33 \quad (Jahre) \end{aligned}$$

Mit Sicherheitspuffer $\Rightarrow x = 8$ *Jahre.*

Für Bauspardarlehen, die ihr Darlehen in verkürzter Regellaufzeit getilgt haben, wird die Indikatorvariable $Y_i = 0$ (Schuldner i fällt nicht aus) gesetzt. Diese Auswahlbedingung lieferte 55.752 endgetilgte Bauspardarlehen im Bestand der Ausgangsbausparkasse.

Tabelle 5.2 zeigt die Verteilung der ausgefallenen und endgetilgten Bauspardarlehen auf Trainings- und Testmenge. Die Menge der ausgefallenen Konten werden wir

	ausgefallen	endgetilgt	$\sum$
Training	1.365	27.876	29.241
Validierung	1.365	27.876	29.241
$\sum$	2.730	55.752	58.482

Tabelle 5.2: Verteilung der ermittelten Daten auf Trainings- und Testmenge

im Folgenden mit G^+, die Menge der endgetilgten Konten mit G^- bezeichnen. Zur Erstellung der formalen Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ wurden verschiedene Stichproben, die jeweils 600 Bausparkonten umfassten, aus den Trainingsmengen G^+ und G^- erstellt. Die weiteren Konten der Trainingsmenge wurden zur Signifikanzprüfung der Implikationen verwendet. Die Testmenge diente ausschließlich zur Überprüfung des Modells.

Vor der Datenkodierung und der Erstellung der Kontexte, werden die Daten hinsichtlich ausgewählter Merkmale analysiert. Diese Analyse bietet die Möglichkeit das ermittelte Regelwerk zusätzlich zu verifizieren bzw. zu bestätigen.

5.5 Darlehensanalyse

Im Folgenden werden ausgewählte Kreditmerkmale der Mengen G^+ und G^- hinsichtlich ihrer relativen Häufigkeit des Auftretens untersucht. Grundlage der Untersuchung war die Menge G^+, sowie eine 10 %-ige Stichprobe der Menge G^-. Tabelle 5.3 zeigt

Weiteres Konto in Zahlungsschwierigkeiten	Y_i=*1* (in %)	Y_i=*0* (in %)
ja	32.83	0.10
nein	67.17	99.90

Tabelle 5.3: Empirische Verteilung des Merkmals „Weiteres Konto in Zahlungsschwierigkeiten

deutlich, dass der Anteil der vorbelasteten Darlehen in der Menge G^+ höher ist als in der Menge G^-. Nahezu ein Drittel aller Konten in der Menge G^+ sind bereits durch ihre Kontoführung bei weiteren Darlehen negativ aufgefallen. Aus Tabelle 5.4 kann

Berufsgruppen	Y_i=*1* (in %)	Y_i=*0* (in %)
Arbeiter	29.61	18.31
Angestellter	43.54	50.13
Beamter	2.91	6.06
Rentner	4.31	10.02
Selbständige	11.93	6.89
Juristische Person	0.11	0.09
ohne Beruf	7.59	8.50

Tabelle 5.4: Empirische Verteilung des Kreditmerkmals „Berufsgruppe”

geschlossen werden, dass die Berufsgruppe für den Ausfall bzw. Nichtausfall eines Darlehens von Bedeutung ist. Der relative Anteil von Arbeitern und Selbständigen ist in der Menge G^+ deutlich höher. Demgegenüber steht ein geringeres Auftreten von Beamten und Rentnern in der Menge G^+. Es ist natürlich möglich, dass die Berufsgruppe nur in Verbindung mit anderen Merkmalen eine signifikante Rolle spielt.

Die Merkmale Spardauer und Alter in Jahren sind aus Gründen der Übersichtlichkeit graphisch dargestellt. Abbildung 5.1 zeigt die empirische Verteilung des Merkmals Spardauer. Dabei kann vermutet werden, dass es sich um keinen linearen Zusammenhang handelt, da die Ausfallwahrscheinlichkeit nicht mit jedem Jahr der Spardauer

steigt. Vielmehr kristallisieren sich zwei Peaks heraus. Die Menge G^- weist einen Höhepunkt bei einer Spardauer von fünf Jahren auf, danach fällt der Anteil im Bestand wieder ab. Hingegen zeigt die Menge G^+ einen Höhepunkt bei einer Spardauer von acht Jahren.

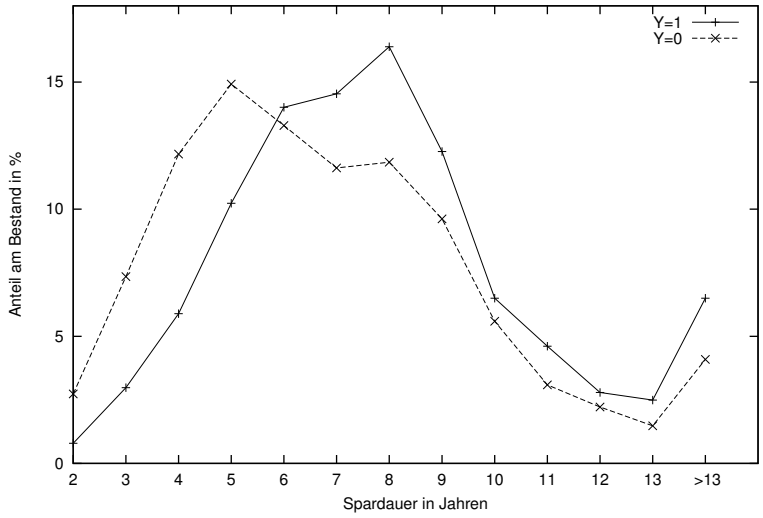


Abbildung 5.1: Graphische Darstellung des Merkmals Spardauer

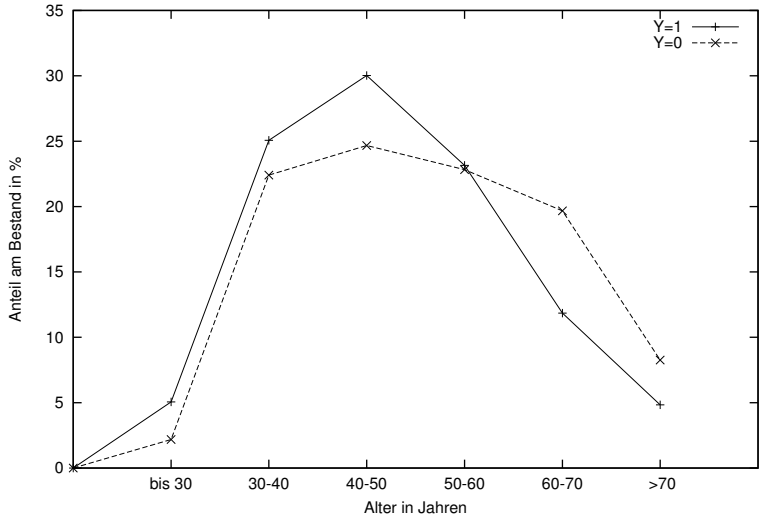


Abbildung 5.2: Graphische Darstellung des Merkmals Alter

In der Menge G^- haben nahezu 40 % der Bausparkonten nach fünf Jahren die Sparphase verlassen und sich das Darlehen auszahlen lassen. Im Gegensatz dazu haben in der Menge G^- nur 20 % bereits nach fünf Jahren die Sparphase verlassen. Betrachtet man die Verteilungsfunktion nach acht Jahren ergibt sich ein ähnliches Bild. Nahe-

zu drei Viertel aller Darlehen in G^- haben die Sparphase nach acht Jahren verlassen, wohingegen der Anteil in der Menge G^+ bei ca. zwei Drittel liegt. Nach neun Jahren Spardauer beginnt sich diese Differenz jedoch zu verringern. Diese Ergebnisse lassen darauf schließen, dass die Verteilung des Merkmals nicht linear ist. Signifikante Unterschiede scheinen hauptsächlich in den ersten acht Jahren zu liegen.
Ähnliche Rückschlüsse erlaubt Abbildung 5.2. Auch hier trifft die Aussage: „mit jedem zusätzlichen Lebensjahr erhöht sich die Ausfallwahrscheinlichkeit" nicht zu. Vielmehr sind die Anteile in der Menge G^+ bis zum 50. Lebensjahr stets höher als die in der Menge G^-. Ab dem 50. Lebensjahr sind die Anteile in Menge G^- allerdings höher.

5.6 Datenaufbereitung

Im folgenden Abschnitt wird auf die Ausprägung und Kodierung der Merkmale eingegangen. Da im Rahmen der Implikationentheorie der formalen Begriffsanalyse ausschließlich Aussagen über vorhandene Merkmale möglich sind, werden z. B. aus dem Merkmal „Weitere Bausparverträge in Zahlungsschwierigkeiten" zwei Variablen erzeugt.

5.6.1 Erläuterungen zu den Merkmalen

- **Weiterer Bausparvertrag vorhanden**
  Da die Anzahl der Bausparverträge nicht beschränkt ist, kann ein Bausparer mehrere Bausparverträge besitzen. Allerdings sind für unsere Untersuchung nur weitere Bausparverträge von Interesse, die sich ebenfalls in der Darlehensphase befinden.

- **Weitere Bausparverträge in Zahlungsschwierigkeiten**
  Es wird ermittelt, ob sich ein weiteres Bauspardarlehen in Zahlungsschwierigkeiten befindet.

- **Weitere Bausparverträge nicht in Zahlungsschwierigkeiten**
  Der Darlehensnehmer besitzt weitere Verträge in der Darlehensphase, von denen keiner in Zahlungsverzug ist.

- **Darlehenshöhe**
  Die maximale Darlehenshöhe (dh_{max}) errechnet sich tarifspezifisch aus dem Produkt von Vertragssumme und 1–Mindestanspargrad, der jeweils 40 % bzw. 50 % beträgt.

$$\begin{aligned} dh_{max} &= bs \cdot (1 - 0.40) \quad \text{bzw.} \\ dh_{max} &= bs \cdot (1 - 0.50) \end{aligned}$$

Der ausgezahlte Darlehensbetrag liegt aber praktisch immer unter dh_{max}, da das Guthaben allein durch Zinsen, die im Zeitraum zwischen positiver Bewertung und Zuteilung anfallen, über den von der Bausparkasse geforderten Mindestanspargrad wächst.

- **Tarife**
  Das Tarifwerk der Ausgangsbausparkasse umfasst eine Vielzahl an Tarifen, die sich in den Tarifbedingungen sowie im Entstehungsjahr unterscheiden. In der Tarifgeneration 2003 (d. h. der Tarif wurde im Jahr 2003 erstmalig angeboten) befand sich zum Untersuchungszeitpunkt kein Darlehen im Bestand, das unseren Bedingungen genügt. Daher beschränkt sich die Untersuchung auf ältere Tarifgenerationen. Generell wird zwischen den Tarifen 1–4 unterschieden. Die Tarife 1 und 3 wurden bei der Untersuchung zusammengelegt, da ihre Konditionen im Darlehensbereich nahezu identisch sind. Für die Untersuchung werden daher drei Haupttarife (Tarif 1/Tarif 3, Tarif 2 und Tarif 4) berücksichtigt. Die unterschiedlichen Mindesttilgungen der Tarife wurden bei der Ermittlung von Zahlungsrückständen und der Regellaufzeit berücksichtigt.

- **Berufe/Berufsgruppen**
  Die Ausgangsbausparkasse erlaubt 26 verschiedene Ausprägungen im Datenfeld Beruf, die zu sieben ähnlichen Berufsgruppen zusammengefasst wurden.

- **Spardauer**
  Die Spardauer wird aus der Differenz des Jahres der Zuteilung und dem Abschlussjahr gebildet. Verschiedene am ZAIK durchgeführte Clusterungen von Bausparverträgen mit abgeschlossener Sparphase ergaben eine Vielzahl unterschiedlicher Sparertypen (z. B. Soforteinzahler, Regelsparer, Nullsparer, . . .), die sich in ihrer Sparleistung und demzufolge auch in der Spardauer unterscheiden.

- **Wohnungsbauprämie (WoP)**
  Der Staat fördert das Bausparen mit der Wohnungsbauprämie. Für die jährlichen Sparbeträge erhalten die Bausparer derzeit[2] 8.8 % p.a. auf einen maximalen Sparbetrag. Allerdings gibt es bestimmte Einkommensgrenzen, die nicht überschritten werden dürfen. Maximale WoP erhalten jene Sparer, welche die Einkommensgrenze nicht überschreiten und jährlich einen konstanten Betrag sparen, der dem maximalen Sparbetrag entspricht.

- **Alter**
  Das Alter des Bausparers wird anhand des Feldes „Geburtsdatum/Jahr" aus den Bauspardaten ermittelt.

[2] Seit dem 01.01.2004 beträgt die Wohnungsbauprämie nur noch 8.8 % auf maximal 512 Euro für Alleinstehende bzw. 1.024 Euro Sparzahlungen für Verheiratete, bei konstanter Einkommensobergrenze.

5.6.2 Datenkodierung

Die Anwendung des verbandstheoretischen Implikationenmodells fordert Daten in binärer Form, d. h. rationale Einträge müssen diskretisiert werden. Die Diskretisierung orientiert sich dabei an der inhaltlichen Bedeutung der Merkmale. Folgende rationale Merkmale müssen diskretisiert werden:

1. **Darlehenshöhe**
   Die Diskretisierung der Variablen Darlehenshöhe orientiert sich an der notwendigen Absicherung des Bauspardarlehens [EFM04]. Für geringe Bauspardarlehen genügt häufig ein vereinfachtes Kreditantragsverfahren, mit zunehmender Darlehenshöhe steigt der Grad der notwendigen Darlehensabsicherung.

2. **Spardauer**
   Die Diskretisierung orientiert sich an der vorgeschlagenen Regelsparrate der Ausgangsbausparkasse. Eine Beispielberechnung für den Tarif 1 findet sich in Unterabschnitt 4.2.2.

3. **Höhe der Wohnungsbauprämie**
   Die Höhe der Wohnungsbauprämie wird in %, bezogen auf das Guthaben des Bausparkontos, gemessen. Dabei wird zwischen geringem und hohem Wohnungsbauprämienanteil am Guthaben unterschieden.

4. **Alter**
   Bei der Diskretisierung der Variablen Alter wurde die aus dem Alter resultierende berufliche Lebenssituation des Bausparers berücksichtigt.

Der Merkmalsvektor $X \in \mathbb{B}^{27}$ besitzt folgende Einträge:

- x_1: Weiterer Bausparvertrag in der Darlehensphase vorhanden? ja/nein
- x_2: Weitere Bausparverträge in Zahlungsschwierigkeiten? ja/nein
- x_3: Weitere Bausparverträge nicht in Zahlungsschwierigkeiten? ja/nein
- x_4: Darlehenshöhe $\leq$ 10.000 Euro? ja/nein
- x_5: 10.000 $\geq$ Darlehenshöhe $\leq$ 15.000 Euro? ja/nein
- x_6: Darlehenshöhe $\geq$ 15.000 Euro? ja/nein
- x_7: Tarif 1/Tarif 3? ja/nein
- x_8: Tarif 2? ja/nein
- x_9: Tarif 4? ja/nein
- x_{10}: Beruf: Arbeiter? ja/nein

- x_{11}: Beruf: Kein Arbeiter? ja/nein
- x_{12}: Beruf: Angestellter? ja/nein
- x_{13}: Beruf: Kein Angestellter? ja/nein
- x_{14}: Beruf: Beamter? ja/nein
- x_{15}: Beruf: Kein Beamter? ja/nein
- x_{16}: Beruf: Rentner? ja/nein
- x_{17}: Beruf: Kein Rentner? ja/nein
- x_{18}: Beruf: Selbstständig? ja/nein
- x_{19}: Beruf: Nicht Selbstständig? ja/nein
- x_{20}: Spardauer $\leq$ 6 Jahre? ja/nein
- x_{21}: 6 Jahr $\geq$ Spardauer $\leq$ 10 Jahre? ja/nein
- x_{22}: Spardauer $\geq$ 10 Jahre? ja/nein
- x_{23}: WoP–Anteil am Guthaben $\leq$ 4% Prozent? ja/nein
- x_{24}: WoP–Anteil am Guthaben $\geq$ 10% Prozent? ja/nein
- x_{25}: Alter: jünger als 40 Jahre? ja/nein
- x_{26}: Alter: zwischen 40 und 60 Jahren? ja/nein
- x_{27}: Alter: über 60 Jahre? ja/nein

Mit Hilfe dieser Kodierung wurden aus den Mengen G^+ und G^- zwei formale Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ erzeugt. In einem weiteren Schritt wurde die Stammbasis der Implikationen für $\mathbb{K}^+$ und $\mathbb{K}^-$ ermittelt. Nach Auswahl der relevanten Implikationen in der Stammbasis, wurde mit einem χ^2–Unabhängigkeitstest deren Signifikanz überprüft. Abschließend wurden die Implikationenmengen $\mathcal{L}_1$ und $\mathcal{L}_2$ anhand des Medians x_{med} zusammengefasst, und die bedingten Wahrscheinlichkeiten für alle Teilmengen empirisch ermittelt.

5.7 Ergebnisse

In den folgenden Abschnitten soll die Klassifikationsgüte des Modells sowie die Übertragbarkeit der Modellierung auf eine weitere Bausparkasse (Validierungsbausparkasse) dargestellt werden. Danach wird auf eine Verwendung der ermittelten Ausfallwahrscheinlichkeiten im Rahmen von IRB–Ansätzen eingegangen. Wir wollen mit dem zusammengefassten Regelwerk beginnen.

5.7.1 Regelwerk

Die Medianzusammenfassung der Implikationenmengen $\mathcal{L}_1$ und $\mathcal{L}_2$ lieferte die folgenden Mengen $\mathcal{L}_{11}$, $\mathcal{L}_{12}$, $\mathcal{L}_{21}$ und $\mathcal{L}_{22}$ in disjunktiver Normalform. Dabei werden die Implikationen so wiedergegeben, wie sie vom Next–Closure–Algorithmus erzeugt wurden.

- $\mathcal{L}_{11}$: Stark signifikante Regeln des Kontextes $\mathbb{K}^+$

  Weiteres Konto in Zahlungsschwierigkeiten
  Selbständig $\rightarrow$ Tarif 1/Tarif 3
  Selbständig $\wedge$ Mittlere Spardauer $\rightarrow$ Geringe WoP
  Selbständig $\wedge$ Mittleres Alter $\rightarrow$ Geringe WoP
  Hohes Darlehen $\wedge$ Selbständig $\rightarrow$ Geringe WoP
  Arbeiter $\wedge$ jung $\rightarrow$ Tarif 1/Tarif 3
  Arbeiter $\wedge$ Geringe WoP$\rightarrow$ Tarif 1/Tarif 3

- $\mathcal{L}_{12}$: Signifikante Regeln des Kontextes $\mathbb{K}^+$

  Kein Angestellter $\wedge$ Geringe WoP $\wedge$ Mittleres Alter $\rightarrow$ Kein Rentner
  Kein Beamter $\wedge$ Geringe WoP $\wedge$ Mittleres Alter $\rightarrow$ Kein Rentner
  Kein Angestellter $\wedge$ jung $\rightarrow$ Kein Beamter
  Selbständig $\wedge$ Geringe Spardauer $\rightarrow$ Geringe WoP
  Hohe Spardauer $\wedge$ Mittleres Alter $\rightarrow$ Geringe WoP
  Arbeiter $\wedge$ Mittlere Spardauer $\rightarrow$ Tarif 1/Tarif 3
  Hohe Spardauer $\wedge$ Mittleres Alter $\rightarrow$ Kein Rentner

- $\mathcal{L}_{21}$: Stark signifikante Regeln des Kontextes $\mathbb{K}^-$

  Weiteres Konto nicht in Zahlungsschwierigkeiten
  Nicht Selbständig $\wedge$ Geringe Spardauer $\wedge$ alt $\rightarrow$ Kein Beamter
  Geringe Spardauer $\wedge$ alt $\rightarrow$ Kleines Darlehen
  Kleines Darlehen $\wedge$ Nicht Selbständig $\wedge$ alt $\rightarrow$ Kein Beamter
  Nicht Selbständig $\wedge$ alt $\rightarrow$ Kein Beamter
  Nicht Selbständig $\wedge$ alt $\rightarrow$ Kein Arbeiter
  Kein Angestellter $\wedge$ Geringe Spardauer $\wedge$ alt $\rightarrow$ Kein Beamter

- $\mathcal{L}_{22}$: Signifikante Regeln des Kontextes $\mathbb{K}^-$

  Kleines Darlehen $\wedge$ Kein Angestellter $\wedge$ alt $\rightarrow$ Kein Beamter
  Kleines Darlehen $\wedge$ Kein Arbeiter $\wedge$ alt $\rightarrow$ Kein Beamter
  Beamter $\rightarrow$ Tarif 1/Tarif 3
  Beamter $\rightarrow$ Geringe WoP
  Rentner $\rightarrow$ Tarif 1/Tarif 3

Kein Angestellter $\wedge$ alt $\rightarrow$ Kein Beamter
Kein Angestellter $\wedge$ Geringe Spardauer $\wedge$ alt $\rightarrow$ Geringe WoP

Die in Abschnitt 5.5 durchgeführte Darlehensanalyse lieferte bereits ähnliche Ergebnisse. Dort fielen vor allem die Merkmale Spardauer, Alter und Berufsgruppe in den Stichproben auf. Während die Berufsgruppe Selbständig eine äußerst hohe Signifikanz für einen Darlehensausfall aufweist, tritt die Berufsgruppe Arbeiter nur in Kombination mit anderen Merkmalen als signifikant für einen Kreditausfall in Erscheinung. Das Merkmal „hohe Spardauer" wird aussschließlich zur Klassifizierung ausgefallener Darlehen verwendet, wohingegen das Merkmal „geringe Spardauer" größtenteils zur Bestimmung von endgetilgten Konten verwendet wird. Tritt es allerdings in Kombination mit der Berufsgruppe Selbständig auf, so dient es auch zur Klassifikation ausgefallener Darlehen.

5.7.2 Ermittelte Kreditausfallwahrscheinlichkeiten und erreichte Konten

Ziel der Modellierung war es, eine möglichst hohe Anzahl von Trainingsbeispielen mit dem Regelwerk zu überdecken. Tabelle 5.5 zeigt die Anteile der überdeckten Konten in Trainings- und Testmenge, die in allen Fällen bei nahezu 90 % liegen. Damit kann die Zielformulierung eines hohen Überdeckungsgrades als erfüllt betrachtet werden.

	Anteil überdeckter Trainingsbeispiele (in %)	Anteil überdeckter Testbeispiele (in %)
Ausgefallene	87.77	90.18
Endgetilgte	88.26	89.94
Gesamtanteil	88.23	89.95

Tabelle 5.5: Anteile überdeckter Trainingsbeispiele durch das Regelwerk in der Trainings- und Testmenge der Ausgangsbausparkasse

Wir bezeichnen mit:

$\mathcal{L}_{1_\bullet}$ = Regeln, die aus dem Kontext $\mathbb{K}^+$ ermittelt wurden,
$\mathcal{L}_{2_\bullet}$ = Regeln, die aus dem Kontext $\mathbb{K}^-$ ermittelt wurden.

Der Ergebnisraum Ω besteht aus allen für die Untersuchung relevanten Konten. Dabei bezeichnet G^+ die Menge der ausgefallenen Trainingsbeispiele und G^- die Menge der endgetilgten Trainingsbeispiele. Die Mengen G^+ und G^- liefern eine vollständige Zerlegung des Ergebnisraumes. Für G^+ und G^- gilt:

$$\begin{aligned} G^+ \cap G^- &= \emptyset \\ G^+ \cup G^- &= \Omega \end{aligned}$$

Ein Ereignis A bezeichnet das Vorliegen einer bestimmten Teilmenge aus $\mathcal{P}(\mathcal{L}_{ij})$. Dann gilt für jede Zerlegung und jedes Ereignis A:

$$P(A) \;=\; \sum_{l} P(G^l) \cdot P(A|G^l) \quad l \in \{+,-\}.$$

Dabei bezeichnet $P(G^l)$, $l \in \{+,-\}$ die a priori Verteilung der ausgefallenen und endgetilgten Darlehen im Bestand. Die Größe $P(A|G^l)$, $l \in \{+,-\}$ entspricht der Wahrscheinlichkeit, dass Ereignis A eintrifft, wenn der Bausparvertrag aus der Menge G^l, $l \in \{+,-\}$ stammt. Daher kann für die Ermittlung der bedingten Wahrscheinlichkeiten die Formel von Bayes verwendet werden. Ist $P(A) > 0$ und gelten die Voraussetzungen der Formel der totalen Wahrscheinlichkeit, so gilt für alle l:

$$P(G^l|A) \;=\; \frac{P(G^l) \cdot P(A|G^l)}{\sum_l P(G^l) \cdot P(A|G^l)} \quad l \in \{+,-\}.$$

Für jedes Bausparkonto wird die bedingte Wahrscheinlichkeit ermittelt, dass das Konto aus der Menge der ausgefallenen bzw. endgetilgten Darlehen stammt. Die empirische Bewertung dieser Teilmengen lieferte die Ergebnisse in Tabelle 5.6. Die bedingten Wahrscheinlichkeiten $P(G^+|A)$ sind in einigen Teilmengen recht gering. Dafür ist die Gesamtverteilung der ausgefallenen und endgetilgten Konten im Bestand der Ausgangsbausparkasse verantwortlich. Die a priori Wahrscheinlichkeit für ein endgetilgtes Darlehenskonto beträgt ca. 95 %, demgegenüber steht eine a priori Wahrscheinlichkeit für ausgefallene Konten von ca. 5 %. Die a posteriori Wahrscheinlichkeit wird daher vom hohen Anteil der endgetilgten Darlehenskonten geprägt. Enthält die Teilmenge jedoch die stark signifikanten Regeln $\mathcal{L}_{11}$, so sind die bedingten Wahrscheinlichkeiten deutlich höher. Allerdings stellt sich die Frage, wie die Klassifikationsgüte dieser Ergebnisse bewertet werden kann. Da die Indikatorvariablen Y_i, $i = \{0,1\}$ aller Datensätze in der Trainings- und Testmenge bekannt sind, kann eine ROC–Analyse mit verschiedenen Schwellwerten durchgeführt werden.

5.7.3 ROC–Graphen und AUC–Werte

Im folgenden Abschnitt werden die ROC–Graphen mit zugehörigen AUC–Werten erzeugt. Zur Erstellung der ROC–Graphen werden Sensitivität und Spezifität für jeden Schwellwert $s \in \mathbb{R}$ ermittelt. Als Schwellwerte werden die bedingten Wahrscheinlichkeiten aus Tabelle 5.6 verwendet. Die Graphen in Abbildung 5.3 und die AUC–Werte in Tabelle 5.7 wurden mittels der in Abschnitt 4.8 vorgestellten Algorithmen berechnet. Tabelle 5.7 zeigt einen sensiblen Bereich in dem gleichzeitig die höchsten Werte für Sensitivität und Spezifität erreicht wurden, d. h. die Anzahl der richtig positiven und richtig negativen Datensätze war dort am höchsten.

Vorhandene Regeln des Ereignisses A	$P(G^+\|A) = \pi(x)$ (in %)	$P(G^-\|A)$ (in %)
$\mathcal{L}_{11}$	63.84	36.16
$\mathcal{L}_{12}$	5.72	94.28
$\mathcal{L}_{21}$	2.90	97.10
$\mathcal{L}_{22}$	1.90	98.10
$\mathcal{L}_{11}\mathcal{L}_{12}$	12.12	87.88
$\mathcal{L}_{11}\mathcal{L}_{21}$	35.72	64.28
$\mathcal{L}_{11}\mathcal{L}_{22}$	14.71	85.29
$\mathcal{L}_{12}\mathcal{L}_{21}$	1.06	98.94
$\mathcal{L}_{12}\mathcal{L}_{22}$	2.33	97.67
$\mathcal{L}_{21}\mathcal{L}_{22}$	1.68	98.32
$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}$	3.02	96.98
$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{22}$	99.99	0.01
$\mathcal{L}_{11}\mathcal{L}_{21}\mathcal{L}_{22}$	9.75	90.25
$\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22}$	1.65	98.35
$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22}$	5.44	94.56

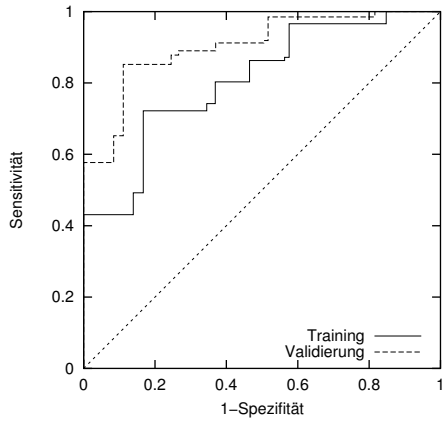
Tabelle 5.6: Bedingte Wahrscheinlichkeiten für einen Kreditausfall im verbandstheoretischen Implikationenmodell

	Sensitivität	Spezifität	AUC–Wert
Trainingsmenge	0.72	0.66	0.735
Testmenge	0.85	0.75	0.852

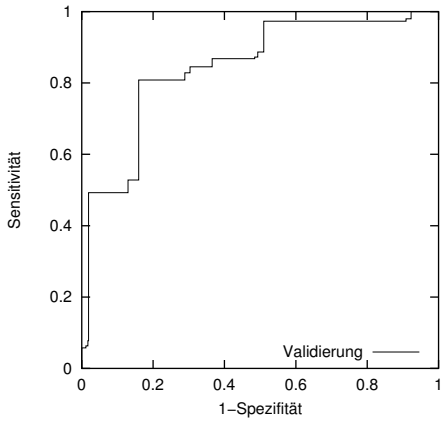
Tabelle 5.7: Sensitivität, Spezifität und AUC–Werte der Trainings- und Testmenge

5.7.4 Zusätzliche Validierung

Zur Überprüfung der Generalisierungsfähigkeit wird das verbandstheoretische Implikationenmodell ohne weitere Anpassungen auf die Originaldaten der Validierungsbausparkasse angewendet. Eine Übertragbarkeit des Modells wäre von Vorteil, da somit eine kosten- und zeitintensive Modellierung in den einzelnen Bausparkassen entfallen könnte. Dazu muss allerdings überprüft werden, inwieweit die Klassifikationsgüte des Modells eine solche Generalisierung zulässt. Die Validierungsbausparkasse besitzt ein ähnliches Tarifwerk sowie ähnliche Tarifbedingungen wie die Ausgangsbausparkasse. Allerdings weist sie eine unterschiedliche Bestandszusammensetzung auf. Zudem existieren lokale Unterschiede in beiden Bausparkassen.
Zur Modellierung wurde vorab das Kollektiv der Validierungsbausparkasse hinsichtlich ausgefallener und endgetilgten Bausparkonten untersucht. Der Untersuchungszeitraum erstreckte sich dabei über die Jahre 1991–2005. Die Kriterien für ausgefallene und endgetilgte Darlehen blieben unverändert. Beide Bausparkassen besitzen iden-



(a) Ausgangsbausparkasse



(b) Validierungsbausparkasse

Abbildung 5.3: ROC–Graphen der Ausgangs- und Validierungsbausparkasse

tische, vertraglich festgelegte Mindesttilgungsraten. Die Analyse lieferte insgesamt 1.369 ausgefallene sowie 69.012 endgetilgte Darlehen im Bestand der Validierungsbausparkasse. Zur Klassifikation der Darlehen wurde das erzeugte bewertete Regelwerk der Ausgangsbausparkasse verwendet. Die Kollektivzusammensetzung der Validierungsbausparkasse ist zwar bekannt, d. h. wir könnten eine kassenspezifische Bewertung durchführen, allerdings soll gerade überprüft werden, inwieweit das verbandstheoretische Modell ohne zusätzliche Analysen auf weitere Bausparkassen übertragen werden kann.

Abbildung 5.3 zeigt die Ergebnisse der Übertragung auf die Validierungsbausparkasse. Für den Flächeninhalt unter der ROC–Kurve ergab sich ein AUC–Wert von 0.7831, der damit zwischen dem Wert der Trainings- und Testmenge der Ausgangsbausparkasse liegt. Die Quote der richtig positiv erkannten Datensätze lag bei 0.81, als richtig negativ klassifiziert wurde ein Anteil von 0.71 Bauspardarlehen. Die Klassifikationsgüte spricht also dafür, dass die Modellierung ohne weitere Anpassungen übertragbar ist.

5.7.5 Zusammenfassung der Ergebnisse

Die verbandstheoretische Modellierung lieferte robuste Klassifikationsergebnisse für die Ausgangs- und Validierungsbausparkasse. Das Modell war mit Hilfe der Bewertung in der Lage, die ausgefallenen von den endgetilgten Bauspardarlehen zu unterscheiden. Zudem konnten überschaubare logische Strukturen aus den Daten abgeleitet werden. Es zeigte sich außerdem, dass die Vermutungen der ersten Darlehensanalyse

aus Abschnitt 5.5 bestätigt werden können. Weiterhin wurde die Annahme bestätigt, dass die Verwendung von vertragsimmanenten Vergangenheitswerten für die Klassifikation sinnvoll ist. Eine weitere Erkenntnis liegt in den gewonnen Merkmalskombinationen. Auch hier konnte die Ausgangsvermutung verifiziert werden, dass viele Merkmale nur in Kombination mit anderen Merkmalen signifikant für die Klassifikation sind.
Zur Klassifikation ungesehener Datensätze ist das Modell aufgrund seiner starken Abhängigkeit von den a priori Wahrscheinlichkeiten aber nur beschränkt anwendbar. Vielmehr kann das Modell durch seine hohe Interpretierbarkeit unterstützend im Entscheidungsprozess mitwirken. Eine kombinierte Anwendung des Modells mit anderen Bewertungsmethoden erscheint daher sinnvoll. Im folgenden Abschnitt wird auf eine weitere interessante Verwendung der ermittelten Ausfallwahrscheinlichkeiten im Rahmen von IRB–Ansätzen eingegangen.

5.8 Verwendung der Ausfallwahrscheinlichkeiten im Rahmen von IRB–Ansätzen

Kreditinstitute müssen im Rahmen von Basel II ihre Kreditnehmer zur Quantifizierung des Kreditrisikos in eine von mindestens acht Bonitätsklassen einordnen. Dabei entfallen sieben Bonitätsklassen auf nicht ausgefallene Darlehen, eine Bonitätsklasse muss für Forderungen, die sich bereits im Verzug befinden, vorgehalten werden. Eine Rating- oder Bonitätsklasse ist definiert als die Einstufung des Schuldnerrisikos auf der Grundlage mehrerer Ratingkriterien, aus denen die Ausfallwahrscheinlichkeit (PD) der Klasse abgeleitet werden kann [Bas04][3]. Die Zuordnung der Kreditnehmer zu den Ratingklassen muss dabei verschiedene Kriterien erfüllen. Sie muss zum einen gewährleisten, dass Kreditnehmer mit ähnlichen Risiken derselben Ratingklasse zugeordnet werden, zum anderen muss die Zuordnung nachvollziehbar und interpretierbar sein [Bas04][4]. Mit Hilfe der Ausfallwahrscheinlichkeiten und zusätzlichen Risikoparametern kann die aufsichtlich geforderte Eigenkapitalunterlegung zur Abdeckung der unerwarteten Verluste ermittelt werden. Die Ergebnisse des verbandstheoretischen Implikationenmodell können zur Bildung solcher Bonitätsklassen verwendet werden. Das Modell zerlegt den Bestand an Bauspardarlehen mit Hilfe logischer Regeln in verschiedene Klassen, die sich durch ihre vorhandenen Merkmalskombinationen definieren. Jeder Kreditnehmer wird anhand seiner vorhandenen Merkmale genau einer Klasse, nämlich der maximalen, zugeordnet. Dadurch wird auf natürliche Weise eine Zuordnung der Kreditnehmer zu den Bonitätsklassen vorgenommen. Das zur Ermittlung der Ausfallwahrscheinlichkeiten verwendete Datenmaterial umfasste einen Zeithorizont von fünf Jahren und genügt somit den Ansprüchen des Baseler Ausschuss für

[3] Vgl. Absatz 405 in [Bas04].
[4] Vgl. Absatz 410 in [Bas04].

Risikoklasse	Vorhandene Regeln	$\pi(x) = PD$ (in %)
1	$\mathcal{L}_{12}\mathcal{L}_{21}$	1.06
2	$\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22}$	1.65
3	$\mathcal{L}_{21}\mathcal{L}_{22}$	1.68
4	$\mathcal{L}_{22}$	1.90
5	$\mathcal{L}_{12}\mathcal{L}_{22}$	2.33
6	$\mathcal{L}_{21}$	2.90
7	$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}$	3.02
8	$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22}$	5.44
9	$\mathcal{L}_{12}$	5.72
10	$\mathcal{L}_{11}\mathcal{L}_{21}\mathcal{L}_{22}$	9.75
11	$\mathcal{L}_{11}\mathcal{L}_{12}$	12.12
12	$\mathcal{L}_{11}\mathcal{L}_{22}$	14.71
13	$\mathcal{L}_{11}\mathcal{L}_{21}$	35.72
14	$\mathcal{L}_{11}$	63.84
15	$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{22}$	99.99

Tabelle 5.8: Einteilung von Risikoklassen mit zugehöriger PD mit Hilfe des verbandstheoretischen Implikationenmodells

Bankenaufsicht[5]. Anhand der Ergebnisse aus Tabelle 5.6 können die in Tabelle 5.8 dargestellten Risikoklassen mit zugehöriger PD gebildet werden. Die Ausfallwahrscheinlichkeit nimmt mit jeder Risikoklasse zu, und damit die Qualität des Darlehens ab. Die Ausfallwahrscheinlichkeiten der Bonitätsklassen erlauben nun die Ermittlung der Risikogewichte und damit die Berechnung des aufsichtlich geforderten Eigenkapitals für einen Kredit. Die aufsichtlich vorgegebenen Risikogewichtsfunktionen sind abhängig von der zugrunde liegenden Forderungsklasse (z. B. Forderungen an Unternehmen und Forderungen an Privatkunden). Im Falle von Forderungen an Privatkunden sind die Risikokomponenten PD und LGD Eingangsparameter für die Risikogewichtsfunktion. Das aufsichtliche Eigenkapital erhält man aus dem Produkt des Risikogewichts (RW) mit der erwarteten Forderungshöhe im Ausfallzeitpunkt (EAD) und der geforderten Eigenkapitalunterlegung (8 %). Ein Berechnungsbeispiel findet sich in Tabelle 2.3 auf Seite 35. Die aufsichtlich vorgegebenen Risikogewichtsfunktionen sind ausführlich in [Bas04, Deu04] dargestellt. Die Risikoklassen müssen allerdings ergänzt werden um eine Klasse für bereits ausgefallene Darlehen sowie eine Klasse zur Bewertung der nicht erfassten Kredite im verbandstheoretischen Implikationenmodell. Da das vorhandene Datenmaterial keine Angaben über zugrunde liegende Sicherheiten beinhaltet, muss der Risikoparameter LGD mit Hilfe anderer Verfahren ermittelt werden, bzw. bei Verwendung des IRB–Basisansatzes wird er aufsichtlich vorgegeben.

[5] Vgl. Absatz 414 in [Bas04].

Der Baseler Ausschuss für Bankenaufsicht stellt in [Bas04] einige Anforderungen an Ratingsysteme zur Ermittlung von Bonitätsklassen und Ausfallwahrscheinlichkeiten. Tabelle 5.9 zeigt beispielhaft wichtige Anforderungen und Kriterien an Ratingverfahren zur Verwendung in IRB–Ansätzen und deren Umsetzung im verbandstheoretischen Implikationenmodell.

Mindestanforderung an ein Ratingverfahren	Umsetzung im verbandstheoretischen Implikationenmodell (v. I.)
Einteilung der Kreditnehmer in mindestens acht Bonitätsklassen [404][6]	Das v. I. erzeugt 15 Klassen von Kreditnehmern mit unterschiedlichen Merkmalsausprägungen, die allerdings noch erweitert werden müssen.
PD–Schätzung [409] für jede Bonitätsklasse	Empirische Ermittlung von bedingten Wahrscheinlichkeiten für jede Klasse.
Schätzung der Risikoparameter LGD, EAD und M [409]	Aufgrund des beschränkten Datenmaterials im v. I. noch nicht berücksichtigt.
Berücksichtigung von Sicherheiten, Garantien und Nachrangigkeiten (erstrangige oder zweitrangige Pfandrechte) [402]	Aufgrund des beschränkten Datenmaterials im v. I. noch nicht berücksichtigt.
Gleichmäßige Verteilung der Kreditnehmer über die Ratingklassen [406]	Wird durch die Zusammenfassung der Implikationen erreicht.
Detaillierte Beschreibung der Risikoklassen, konsistente Zuordnung der Risiken zu den jeweiligen Klassen [410]	Konsistente Zuordnung der Kreditnehmer durch Abfrage der logischen Regeln, welche durch die enthaltenen Implikationen detailliert beschrieben sind.
Detaillierte, nachvollziehbare und interpretierbare Ratingdefinition [410]	Das logische Regelwerk im v. I. ist sehr gut interpretierbar.
Kriterien müssen mit bankinternen Kreditvergaberichtlinien übereinstimmen [410]	Die inhaltliche Diskretisierung der rationalen Merkmale erlaubt eine Einbeziehung interner Verfahren.
Zeithorizont für die PD–Schätzung von mindestens einem Jahr [414]	Empirische Ermittlung der PD auf Grundlage von fünf Jahren.

Tabelle 5.9: Ausgewählte Mindestanforderungen und Kriterien an Ratingverfahren zur Verwendung in IRB–Ansätzen und deren Umsetzung im verbandstheoretischen Implikationenmodell

[6] Die Angaben in den eckigen Klammern beziehen sich auf die zugehörigen Absätze in [Bas04].

Kapitel 6

Ergebnisse der vorgestellten Klassifikationsmodelle

6.1 Neuronale Netze

6.1.1 Untersuchungsaufbau und Datenvorbereitung

Zur Klassifizierung der Darlehenskonten mit Hilfe von neuronalen Netzen wurden aus den ermittelten Daten fünf unterschiedliche Stichproben erzeugt. Diese Stichproben können als repräsentativ angenommen werden, da sie es möglich machen, für unsere inhaltlich abgegrenzte Fragestellung stabile und generalisierbare Analyseergebnisse zu erzielen. Sie genügen dabei folgenden Anforderungen:

- großer Stichprobenumfang
- zufällige Auswahl der Bauspardarlehen
- umfangreicher Betrachtungshorizont 2002–2004
- ausschließliche Betrachtung von Bauspardarlehen

Stichprobe(n)	endgetilgte Darlehen	ausgefallene Darlehen	$\sum$
S 1–4	2.648	2.648	5.296
S 5	5.282	2.648	7.930

Tabelle 6.1: Stichproben zur Modellierung der Ausfallwahrscheinlichkeiten mit Hilfe von neuronalen Netzen

Da zur Modellierung die logistische Aktivierungsfunktion aus Abschnitt 3.1 verwendet wird, wurden die rationalen Merkmalen (Darlehenshöhe, Spardauer, WoP–Anteil und Alter) logarithmiert und anschließend linear so transformiert, dass sie Werte im Intervall $[0, 1]$ annehmen. Zur weiteren Verarbeitung der Daten wurden die fünf Stichproben in eine Trainings-, Cross- und Testmenge zerlegt. Dabei bezeichnet beispielsweise

Stichprobe(n)	Trainingsmenge	Crossmenge	Testmenge	$\sum$
S 1–4	800:800	800:800	1.048:1.048	5.296
S 5	800:1.600	800:1.600	1.048:2.082	7.930

Tabelle 6.2: Zerlegung der Stichproben in Trainings-, Cross- und Testmenge

800:800 das Mischungsverhältnis der endgetilgten und ausgefallenen Darlehenskonten. Die Stichproben S 1–4 wurden stets im Verhältnis 1:1 zerlegt, die Stichprobe S5 wurde im Verhältnis 1:2 aufgeteilt. Alle hier vorgestellten Modellierungen wurden mit dem Programm SNNS (Stuttgart Neural Network Simulator) durchgeführt[1].

6.1.2 Ergebnisse des Pruning–Algorithmus

Bisherige Untersuchungen von neuronalen Netzen bei finanzwirtschaftlichen Problemstellungen [Zim94] haben gezeigt, dass Netze mit einer verdeckten Neuronenschicht völlig ausreichend sind. Daher konzentrieren sich die Untersuchungen auf dreilagige Feed–Forward–Netze.
Getestet wurden Basisnetzwerke mit 17 Inputknoten (Merkmalen), x verdeckten Neuronen und einem Ausgabeneuron. Diese werden im Folgenden kurz mit 17/x/1 bezeichnet. Dabei wurde eine Homogenität der Klassifikationsquoten in Trainings- und Crossmenge angestrebt. Der α– und β–Fehler durfte in beiden Mengen nicht größer als 30 % sein. Beim α–Fehler wird ein Kredit als „gut" beurteilt, obwohl er tatsächlich ausgefallen ist (False Negatives). Hingegen wird beim β–Fehler ein Darlehen als ausgefallen klassifiziert, obwohl das Darlehen vom Kreditnehmer in Regelzeit getilgt wurde (False Positives). Im Sinne der ROC–Analyse entspricht der α–Fehler = 1 – true positive rate = 1 – Sensitivität, der β–Fehler entspricht der false positive rate. Die im Folgenden erwähnte Gesamtklassifikationsquote entspricht der accuracy. Beim α–Fehler entstehen also Kosten durch Zahlungsausfälle und zusätzlichen Verwaltungsaufwand. Der β–Fehler verursacht Opportunitätskosten, da dem Kreditinstitut Zinserträge und Provisionen entgangen sind. In der Regel kommt der Minimierung des α–Fehler größere Bedeutung zu, da die Kosten für Zahlungsausfälle die Opportunitätskosten überschreiten.

Tabelle 6.3 zeigt die Klassifikationsergebnisse der verschiedenen Basisnetzwerke. Die Bandbreiten in Tabelle 6.3 geben Ergebnisse an, die aufgrund verschiedener Stichprobenzusammensetzungen (S 1–5) erzielt wurden. Allen Basisnetzwerken war gemeinsam, dass in den Auto–Pruning Läufen eine recht hohe Klassifikationsqualität erreicht wurde. Weiterhin zeigt Tabelle 6.3 dass Netze mit zehn verdeckten Neuronen anscheinend überdimensioniert waren. Gleiche Klassifikationsqualitäten lassen sich auch

[1] Das Programm ist unter `http://www-ra.informatik.uni-tuebingen.de/SNNS/` frei verfügbar.

Testmenge Gesamt-klassifikations-niveau	Struktur des Basis-netzwerkes	Anfangs-gewichte	End-gewichte	Anzahl benutzter verdeckte knoten	Anzahl noch benutzer Inputknoten (Merkmale)
$\sim$ **80** %	17/10/1	170	18–32	4–6	9–15
	17/8/1	136	17–38	3–7	11–14
	17/6/1	102	20–38	4–6	10–16
	17/5/1	85	24–37	4–5	14–15
	17/4/1	68	16–30	3–4	12–13
	17/3/1	51	20–24	3	14
$\sim$ **77** %	17/10/1	170	17–34	4–6	12–15
	17/8/1	136	36	6	15
	17/6/1	102	17–21	4–5	11–12
	17/5/1	85	26–28	4–5	12–14
	17/4/1	68	21–30	3–4	13–16

Tabelle 6.3: Klassifikationsergebnisse bei unterschiedlichen Basisnetzwerken

mit weniger Gewichten und einer geringeren Anzahl verdeckter Neuronen erreichen.

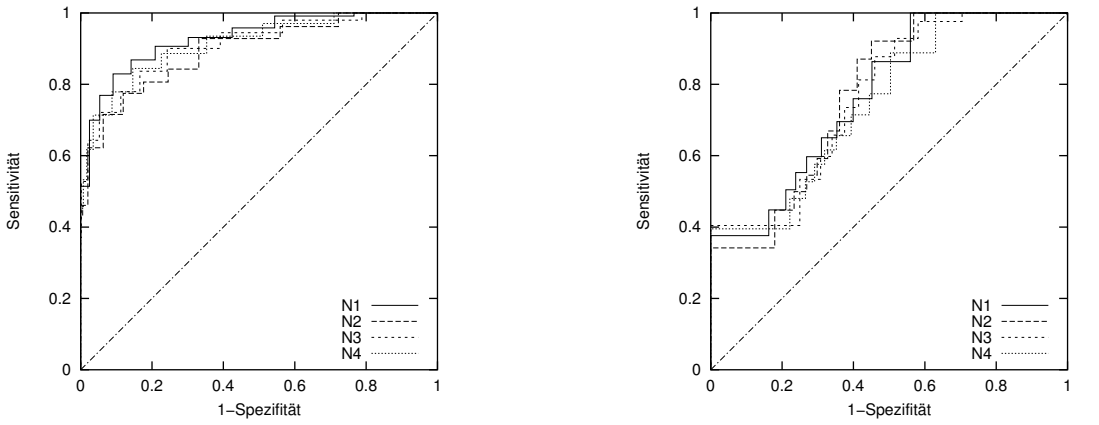
Allerdings zeigte sich auch, dass Abweichungen vom Verhältnis 1:1 in den Stichproben die Ergebnisse deutlich verschlechtert. Tabelle 6.4 zeigt, dass bei einer Abweichung zwar die Gesamtklassifikationsquote nahezu erhalten bleibt, der α–Fehler sich aber mehr als verdoppelt. Das neuronale Netzwerk hat die endgetilgten Darlehen besonders gut gelernt, und neigt daher dazu, Darlehen als endgetilgt einzustufen. Dies zeigt sich auch darin, dass der β–Fehler sich mehr als halbiert hat. Um die angestrebte Homogenität in α– und β–Fehler zu erhalten, werden für die folgenden Untersuchungen ausschließlich Stichproben im Verhältnis 1:1 verwendet.

Anzahl lebender Gewichte	Proportion Trainings- und Crossmenge	Testmenge α–Fehler (in %)	Testmenge β–Fehler (in %)	Testmenge accuracy (in %)
24	1:1	21.8	17.0	80.7
24	1:2	42.5	8.3	80.3
24	1:1	13.2	20.9	82.9
21	1:2	39.6	6.2	82.6

Tabelle 6.4: Klassifikationsergebnisse bei unterschiedlichen Proportionen in den Lernmengen

6.1.3 Feinoptimierung der Ergebnisse

Zur Feinoptimierung wurden aus den Ergebnissen in Tabelle 6.3 die Netze ausgewählt, die eine besonders hohe Klassifikationsgüte bei möglichst ausgeglichenem α– und β–Fehler in der Crossmenge aufwiesen und eine minimale Netzwerkstruktur besitzen. Diese wurden mit manuellem Training und verschiedenen Lernraten optimiert. Die Feinoptimierung lieferte allerdings nur sehr geringe Verbesserung wie Tabelle 6.5 zeigt. Dabei bezeichnet der Index S_j die Stichprobe, aus der das Netz stammt.
Die neuronalen Netze ermitteln für jedes Darlehen eine Ausfallwahrscheinlichkeit $\pi(x)$. Dies wirft die Frage auf, ab welcher Wahrscheinlichkeit $\pi(x)$ ein Darlehen als ausgefallen bzw. endgetilgt zu bewerten ist. Zur Analyse der Ergebnisse werden für die vier Endnetze die zugehörigen ROC–Graphen erstellt. Abbildung 6.1 zeigt die ROC–Graphen der vier neuronalen Netze aus der Feinoptimierung. Für den Flächeninhalt unter den vier Netzen ergaben sich die in Tabelle 6.6 dargestellten Werte für die Ausgangsbausparkasse.



(a) Ausgangsbausparkasse

(b) Validierungsbausparkasse

Abbildung 6.1: ROC–Graphen der vier neuronalen Netze der Ausgangs- und Validierungsbausparkasse

Mit der Festlegung eines Basisnetzwerkes und der Ermittlung der minimalen Netzwerkstruktur mit Hilfe eines Pruning–Algorithmus war die Gesamtklassifikationsquote nahezu festgelegt. Mit manuellem Training konnten die Ergebnisse nur minimal verbessert werden.
Auffallend war, dass eine Abweichung der Zusammensetzung der Stichproben vom Verhältnis 1:1 zu einer drastischen Verschlechterung des Ergebnisses führte. Dagegen spielt die Stichprobenzusammensetzung eine untergeordnete Rolle, da die vier ausgewählten Endnetze aus drei verschiedenen Stichproben stammten und alle Ergebnis-

Netz	Anzahl Input-knoten	Anzahl verdeckter Knoten	Gewichte	Cross-menge α–Fehler (in %)	Cross-menge β–Fehler (in %)	Testmenge accuracy (in %)
$N1_{S_1}$	15	4	24	27.5	18.8	82.9
$N2_{S_2}$	11	3	17	23.5	17.8	79.9
$N3_{S_3}$	10	4	20	23.6	18.0	80.7
$N4_{S_3}$	14	3	24	26.1	16.1	81.6

Tabelle 6.5: Ergebnisse der Feinoptimierung bei vier ausgewählten Netzen

se zwischen 83 % und 80 % Gesamtklassifikationsquote lieferten. Die erfolgreichsten Netze wiesen alle drei bis vier verdeckte Knoten auf. Netzwerke mit einer höheren Anzahl an verdeckten Neuronen scheinen überdimensioniert zu sein, was die Aussagen in [Zim94] bestätigt.

6.1.4 Zusätzliche Validierung

Die ausgewählten realen Datensätze der Validierungsbausparkasse wurden mit Hilfe der vier neuronalen Netze klassifiziert. Die Netze wurden unverändert zur Ermittlung der Ausfallwahrscheinlichkeiten $\pi(x)$ verwendet. Es erfolgte kein neues Training. Betrachtet man beide ROC–Graphen, so fällt auf, dass das Klassifikationsniveau bei

Netz	AUC–Wert
$N1_{S1}$	0.9064
$N2_{S2}$	0.8646
$N3_{S3}$	0.8844
$N4_{S3}$	0.8893

Netz	AUC–Wert
$N1_{S1}$	0.6974
$N2_{S2}$	0.7111
$N3_{S3}$	0.6962
$N4_{S3}$	0.6706

Tabelle 6.6: Von links nach rechts: AUC–Werte der vier neuronalen Netze für die Ausgangs- und Validierungsbausparkasse

der Validierungsbausparkasse deutlich gesunken ist. Der Graph besitzt auf der y–Achse deutlich geringere Werte, was einer gesunkenen Sensitivität entspricht. Dies lässt vermuten, dass es sich hier um einen Overfitting–Effekt handelt, da die Netze speziell auf die Ausgangsbausparkasse trainiert wurden. Eine nähere Analyse dieses Effektes findet sich in Kapitel 7. Für den Flächeninhalt unter den vier Netzen ergaben sich die in Tabelle 6.6 angegebenen Werte.

6.2 Modell von Truemper

6.2.1 Untersuchungsaufbau und Datenvorbereitung

Zur Klassifizierung der Kreditausfälle mit dem logischen Modell von Truemper wurden zehn verschiedene Stichproben aus den Daten erzeugt. Diese unterscheiden sich zum einen in der Anzahl der Datensätze und zum anderen im Verhältnis der Zusammensetzung der Stichproben. Die Stichprobenzusammensetzung ist in Tabelle 6.7 abgebildet. Sowohl die Diskretisierung der Eingangsdaten, als auch die anschließende Klassifizierung wurden mit dem LEIBNIZ–System durchgeführt.

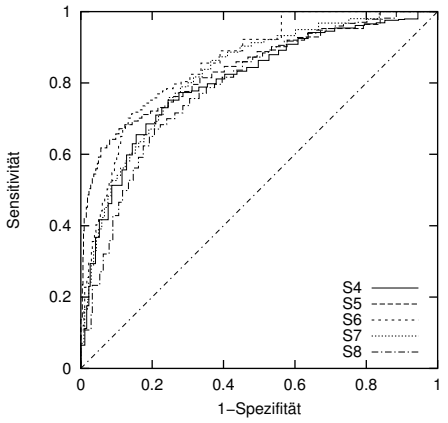
Stichproben-bezeichnung	Verhältnis aus:end	Gesamtanzahl Darlehen
S1–S8	1:1	50 – 1.000
S9–S10	1:2	150 – 300

Tabelle 6.7: Trainingsstichproben für das logische Modell von Truemper

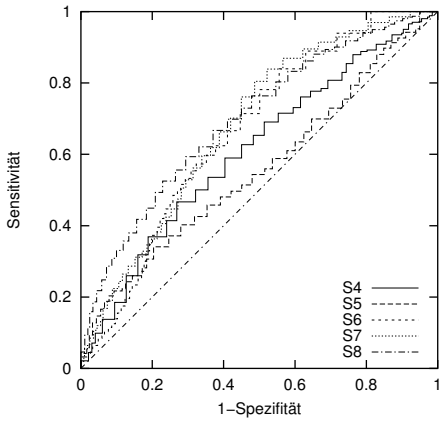
Zur Validierung wurde eine Testmenge mit 5.000 Bauspardarlehen erstellt. Dies entspricht gerade der maximalen Anzahl von Datensätzen die das LEIBNIZ–System verarbeiten kann. Alle zehn erzeugten Klassifikatoren wurden mit Hilfe des Testdatensatzes validiert. Die folgenden Vergleiche der Testergebnisse basieren alle auf demselben Testdatensatz. Zur Diskretisierung der rationalen Variablen Darlehenshöhe, Spardauer, Alter und WoP–Anteil wurde das „Cutpoint–Verfahren" verwendet, das im Programmpaket „Cutcc" des LEIBNIZ–System enthalten ist.

6.2.2 Ergebnisse und zusätzliche Validierung

Ähnlich wie bei der Modellierung mit neuronalen Netzen wiesen die beiden Stichproben, die vom Verteilungsverhältnis 1:1 abwichen, relativ schlechte Ergebnisse auf. Die Stichproben, die weniger als 300 Bausparkonten umfassten, fielen ebenso durch eine schlechtere Klassifikationsgüte auf. Zur näheren Auswertung wurden daher die Stichproben S4–S8 verwendet. Das LEIBNIZ–System erzeugt 40 logische DNF–Formeln mit unterschiedlichen DNF–Klauseln. Eine logische DNF–Formel aus dem Modell von Truemper ist beispielhaft in Abbildung 6.3 dargestellt.
Mit Hilfe der DNF–Formeln kann eine Bewertung der Datensätze vorgenommen werden, indem die Anzahl der Stimmen für jeden Datensatz addiert werden. Dabei entspricht jede erfüllte DNF–Formel genau einer Stimme. Zur Ermittlung von Sensitivität und Spezifität wurde die vom LEIBNIZ–System ermittelte Stimmanzahl der Testmenge verwendet.
Mit Hilfe des erweiterten Algorithmus 1 auf Seite 82 wurde daraus eine ROC–Kurve



(a) Ausgangsbausparkasse



(b) Validierungsbausparkasse

Abbildung 6.2: Klassifikationsergebnisse des Modells von Truemper für die Ausgangs- und Validierungsbausparkasse

Stichprobe	AUC–Wert
S4	0.7953
S5	0.8325
S6	0.8340
S7	0.8185
S8	0.7793

Stichprobe	AUC–Wert
S4	0.5939
S5	0.5450
S6	0.6441
S7	0.6691
S8	0.6853

Tabelle 6.8: Von links nach rechts: AUC–Werte des Modells von Truemper für die Stichproben der Ausgangs- und Validierungsbausparkasse

erzeugt, die in Abbildung 6.2 dargestellt ist. Die zugehörigen AUC–Werte sind in Tabelle 6.8 dargestellt.
Zur Überprüfung der Übertragungsfähigkeit wird eine Testmenge aus dem realen Kollektiv der Validierungsbausparkasse verwendet, um die erzeugten Regelwerke des Modells von Truemper ohne weitere Anpassungen zu testen. Dazu werden ebenfalls die Stichproben S4–S8 verwendet. Die Ergebnisse dieser Generalisierung sind in Abbildung 6.2 und Tabelle 6.8 dargestellt. Dabei zeigt sich, dass sich die ROC–Graphen sowie die AUC–Werte deutlich verschlechtert haben. Sie liegen bei der Validierungsbausparkasse nur knapp über der eingezeichneten Diagonalen $y = x$, die dem „bloßen Raten" entspricht.

```
Formula 1 B 14 min (True means vote for B, 14 min size clause(s))
clause size literals
  1   7     3   7  10  28 -31 -33 -36
  2   6     3 -11  20  22  28 -36
  3   8     3   7 -18  22  26  28 -31 -35
  4   5     3  -9 -18  26  29
  5   5     1  20  22  31  37
  6   5     6 -18  28 -34  36
  7   2    11  26
  8   4     3   9  29 -37
  9   3     8  22  37
 10   3     6  22 -32
 11   4     6   7  30  37
 12   4     3  14  33 -37
 13   3     5  18  30
 14   3     6   9  36
       Logic notation
       [ WK_S_1 & Darlehenshoehe_1 & Tarifklasse_1
        & Kein_Selbst_2 & -Spardauer_3 & -WOP_2
        & -Alter_2 ] |
       [ WK_S_1 & -Tarifklasse_2 & Kein_Arbeiter_2
        & Kein_Angestellter_2 & Kein_Selbst_2
        & -Alter_2 ] |
       [ WK_S_1 & Darlehenshoehe_1 & -Beruf_6
        & Kein_Angestellter_2 & Kein_Rentner_2
        & Kein_Selbst_2 & -Spardauer_3 & -Alter_1 ] |
       [ WK_S_1 & -Darlehenshoehe_3 & -Beruf_6
        & Kein_Rentner_2 & Spardauer_1 ] |
       [ WK_1 & Kein_Arbeiter_2 & Kein_Angestellter_2
        & Spardauer_3 & Alter_3 ] |
       [ WK_G_2 & -Beruf_6 & Kein_Selbst_2 & -WOP_3
        & Alter_2 ] |
       [ Tarifklasse_2 & Kein_Rentner_2 ] |
       [ WK_S_1 & Darlehenshoehe_3 & Spardauer_1
        & -Alter_3 ] |
       [ Darlehenshoehe_2 & Kein_Angestellter_2
        & Alter_3 ] |
       [ WK_G_2 & Kein_Angestellter_2 & -WOP_1 ] |
       [ WK_G_2 & Darlehenshoehe_1 & Spardauer_2
        & Alter_3 ] |
       [ WK_S_1 & Beruf_2 & WOP_2 & -Alter_3 ] |
       [ WK_G_1 & Beruf_6 & Spardauer_2 ] |
       [ WK_G_2 & Darlehenshoehe_3 & Alter_2 ]
```

Abbildung 6.3: Beispiel einer logischen DNF–Formel aus dem Modell von Truemper

6.3 Entscheidungsbäume

6.3.1 Untersuchungsaufbau und Datenvorbereitung

Zur Modellierung mit Entscheidungsbäumen wurden 25 verschiedene Trainingsstichproben erstellt. Wie in den vorher untersuchten Modellen unterscheiden sich die Stichproben in der Anzahl der Darlehen und im Verteilungsverhältnis. Die Anzahl der Konten in den Stichproben betrug minimal 100 Konten und maximal 2.730 Konten. Die Verteilungsverhältnisse betrugen 1:1 und 1:2. Im zweiten Fall wurde die Anzahl der endgetilgten Konten verdoppelt. Zur Validierung wurde eine gemeinsame Testmenge aus dem realen Kollektiv der Validierungsbausparkasse erstellt, die 27.241 Konten umfasste. Die Klassifikation wurde mit der Software Weka 3.4 durchgeführt[2]. Zur Diskretisierung der Daten wurde ein entropiebasierter Ansatz verwendet, der den Informationsgehalt eines Merkmals widerspiegelt.

6.3.2 Ergebnisse und zusätzliche Validierung

Bei den Entscheidungsbäumen handelt es sich um einen diskreten Klassifikator. Während bei den vorigen Modellen stets eine ROC–Kurve mit Hilfe von Schwellwerten erstellt werden konnte, wird bei den Entscheidungsbäumen ein diskretes Klassifikationsergebnis ausgegeben. Daher kann auch kein AUC–Wert angegeben werden.
Die ermittelten Entscheidungsbäume unterscheiden sich vor allem hinsichtlich der Anzahl der Blätter und Knoten. Die Stichproben S3 und S4 besitzen nahezu 50 Entscheidungsknoten, was bei der Testmenge zu einem Overfitting–Effekt führt. Die Trainingsstichprobe S1, die am wenigsten Knoten erzeugte, lieferte hingegen die stabilsten Validierungsergebnisse. Abbildung 6.5 zeigt einen Entscheidungsbaum zur Klassifizierung des Kreditrisikos der Stichprobe S5.

Stichprobe	Sensitivität	Spezifität	Anzahl Blätter	Anzahl Knoten
S1	0.827	0.736	18	20
S2	0.839	0.700	15	24
S3	0.786	0.711	28	45
S4	0.793	0.716	28	46
S5	0.797	0.761	16	26

Tabelle 6.9: Klassifikationsergebnisse von fünf ausgewählten Stichproben mit zugehöriger Anzahl von Blättern und Knoten

In einem weiteren Schritt wurden die ermittelten Bäume dazu verwendet, die Testmenge zu klassifizieren. Aus Gründen der Vergleichbarkeit sind in Abbildung 6.4 die

[2]Diese kann kostenlos unter `http://www.cs.waikato.ac.nz/ml/weka/` heruntergeladen werden.

Klassifikationsergebnisse beider Bausparkassen dargestellt. Die Klassifikationsergebnisse der Validierungsbausparkasse liegen im ROC–Graph auf der x–Achse weiter rechts, was einer höheren false positive rate entspricht.

Um einen fairen Vergleich zwischen den verschiedenen Klassifikationsmodellen und dem verbandstheoretischen Implikationenmodell herstellen zu können, wird im folgenden Kapitel für alle Klassifikationsmodelle eine identische Diskretisierung gewählt. Danach findet ein quantitativer und qualitativer Vergleich der Klassifikationsergebnisse statt.

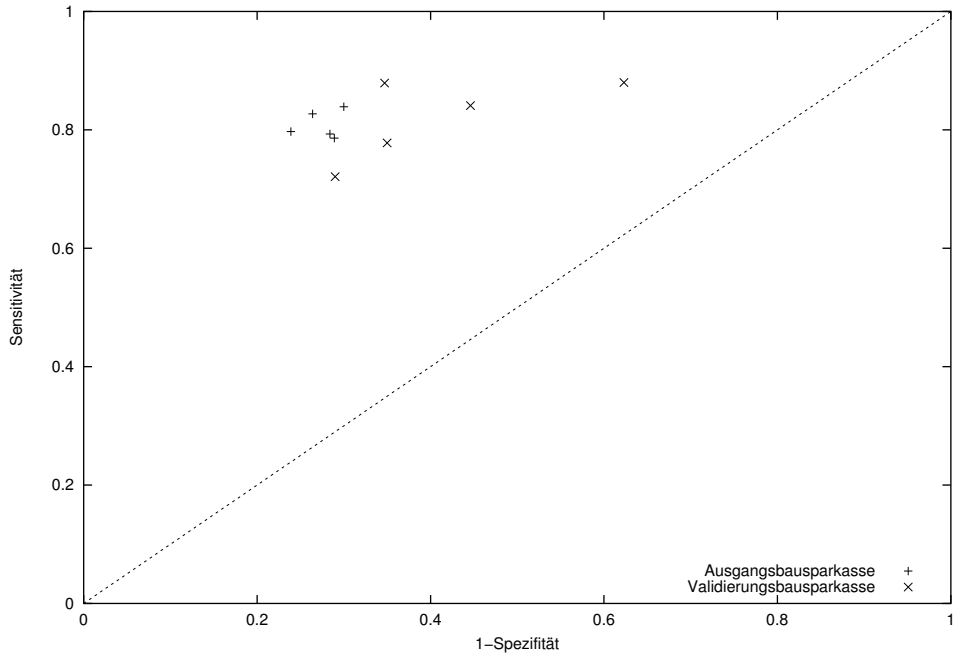


Abbildung 6.4: Klassifikationsergebnisse der Entscheidungsbäume für die Ausgangs- und Validierungsbausparkasse

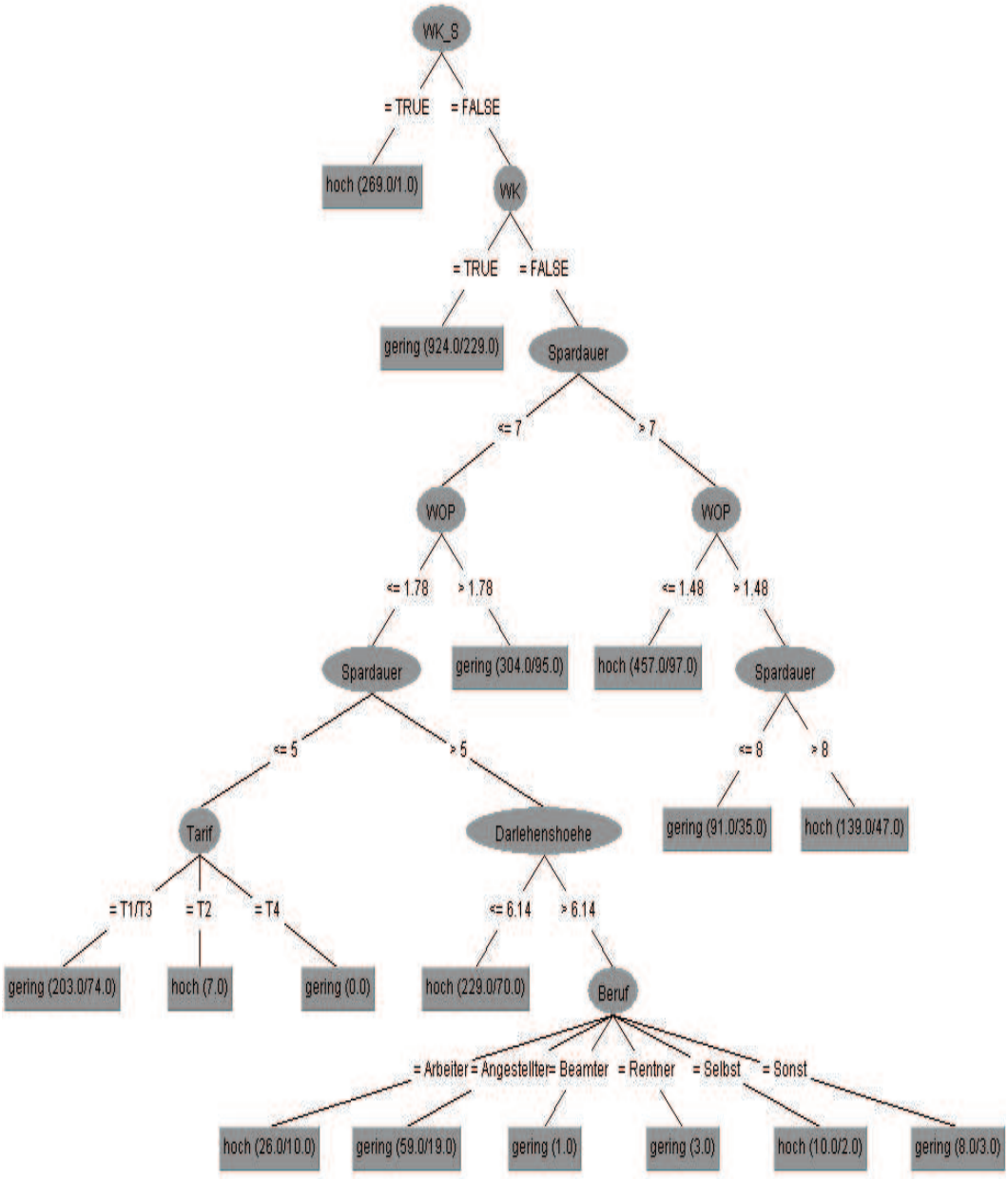


Abbildung 6.5: Entscheidungsbaum zur Klassifizierung des Ausfallrisikos (Stichprobe S5, Ausgangsbausparkasse)

Kapitel 7

Quantitativer und qualitativer Vergleich aller Modelle

7.1 Quantitativer Vergleich

Zum Vergleich der Klassifikationsgüte aller vorgestellten Modelle, werden die Ergebnisse der jeweiligen Testmengen angegeben. Um der Vergleichbarkeit der Modelle Rechnung zu tragen, wurde für die logischen Modelle (Modell von Truemper und Entscheidungsbäume) die inhaltliche Diskretisierung der rationalen Daten gewählt. In einem ersten Schritt werden jeweils die Ergebnisse der Ausgangsbausparkasse verglichen; eine weitere Verifizierung erfolgt mit den Ergebnissen der Validierungsbausparkasse. Abschließend erfolgt eine Interpretation der Ergebnisse.

7.1.1 Neuronale Netze

7.1.1.1 Datenvorbereitung

Die rationalen Eingangsvariablen müssen bei der Verwendung von neuronalen Netzen nicht in logischer Form vorliegen. Sie wurden logarithmiert und für die Verwendung in neuronalen Netzen linear transformiert. Zum Vergleich der Ergebnisse werden die Kenngrößen Sensitivität, Spezifität und der AUC–Wert verwendet. Beim Ergebnisvergleich und der Interpretation der Ergebnisse muss natürlich die unterschiedliche Form der rationalen Eingangsvariablen berücksichtigt werden.

7.1.1.2 Klassifikationsergebnisse der Ausgangsbausparkasse

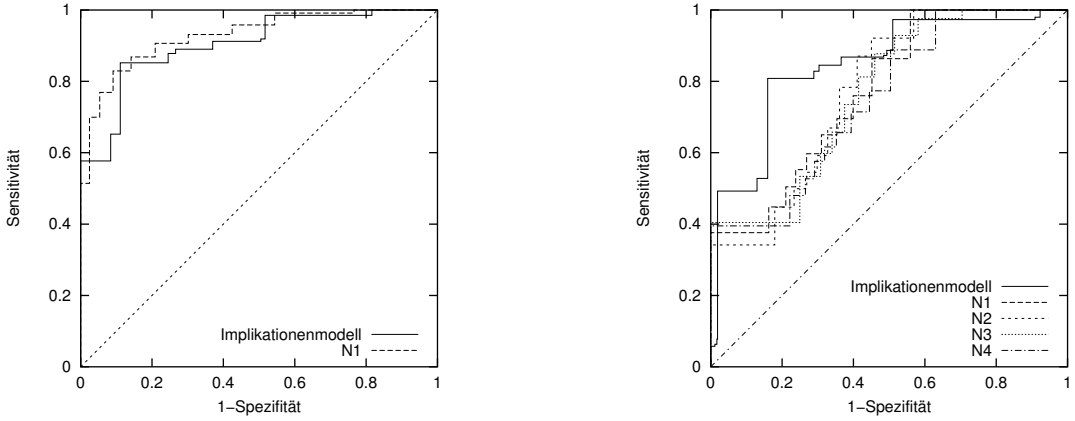
In einem ersten Vergleich werden die Ergebnisse des verbandstheoretischen Implikationenmodells mit den vier neuronalen Netzen verglichen. Abbildung 7.1 zeigt beispielhaft das Ergebnis des neuronalen Netzes N1 im Vergleich mit dem verbandstheoretischen Implikationenmodell. Dabei zeigt sich, dass die neuronalen Netze eine sehr hohe Klassifikationsgüte liefern. Das Niveau des Netzes N1 liegt deutlich über

dem Klassifikationsniveau des verbandstheoretischen Implikationenmodells. Sensitivität und Spezifität liegen im Netz N1 deutlich über 0.8 wie Tabelle 7.1 zeigt. Damit liefern die neuronalen Netze die besten Klassifikationsergebnisse für die Ausgangsbausparkasse.

Sensitivität	Spezifität
0.8683	0.7910
0.8292	0.8588
0.7691	0.9094

Sensitivität	Spezifität
0.5972	0.6898
0.5526	0.7313
0.5044	0.7619

Tabelle 7.1: Von links nach rechts: Sensitivität und Spezifität des neuronalen Netzes N1 der Ausgangs- und Validierungsbausparkasse



(a) Ausgangsbausparkasse

(b) Validierungsbausparkasse

Abbildung 7.1: Vergleich der Klassifikationsergebnisse der Modellierung mit neuronalen Netzen und dem verbandstheoretischen Implikationenmodell der Ausgangs- und Validierungsbausparkasse

7.1.1.3 Klassifikationsergebnisse der Validierungsbausparkasse

In einem weiteren Schritt wird die Übertragbarkeit der neuronalen Netze auf eine weitere Bausparkasse untersucht. Dabei fällt auf, dass bei einer Generalisierung der Netze, die Klassifikationsergebnisse deutlich absinken. Die neuronalen Netze lieferten hier, im Gegensatz zur Ausgangsbausparkasse, die schlechtesten Ergebnisse aller untersuchten Modelle. In Tabelle 7.1 und 7.2 sind Sensitivität, Spezifität und AUC–Werte des Netzes N1 dargestellt, Abbildung 7.1 zeigt die graphische Darstellung der zugehörigen ROC–Graphen.

Modell	Ausgangsbausparkasse	Validierungsbausparkasse
Neuronales Netz[1]	0.9064	0.6974
Verbandstheoretisches Implikationenmodell	0.8523	0.7831

Tabelle 7.2: AUC–Werte des verbandstheoretischen Implikationenmodells und der neuronalen Netze im Vergleich

7.1.2 Modell von Truemper

7.1.2.1 Datenvorbereitung

In Truempers Ansatz der logischen Klassifikation ist die Methode „Cutpoint" zur Diskretisierung von rationalen Daten implementiert, die einen binären Datensatz erzeugt. Um einen realistischen Vergleich herstellen zu können, wird die inhaltliche Diskretisierung auch für das Modell von Truemper verwendet. Dieser Sachverhalt wird im Folgenden durch die Bezeichnung modifiziertes Modell von Truemper umschrieben. Es wurden dieselben Stichproben wie in Unterabschnitt 6.2.1 verwendet. Zum Vergleich der Diskretisierungsverfahren sind in Tabelle 7.3 die „Cutpoints"' einer Trainingsstichprobe dargestellt, die z. B. beim Merkmal Spardauer hohe Ähnlichkeiten zur inhaltlichen Diskretisierung aufweisen. Da im LEIBNIZ–System eine maximale An-

Merkmal	Anzahl „Cutpoints"	1	2	3	4	5	6	7	8
Darlehenshöhe	8	2.94	3.64	4.86	5.46	6.27	6.91	7.83	16.12
Spardauer	5	5.5	6.5	7.5	8.5	10.5			
WoP–Anteil	8	0.02	0.11	0.26	0.47	1.49	2.31	3.01	4.57
Alter	8	30.5	36.5	37.5	41.5	42.5	43.5	53.5	65.5

Tabelle 7.3: „Cutpoints" der Trainingsmenge S7 der Ausgangsbausparkasse

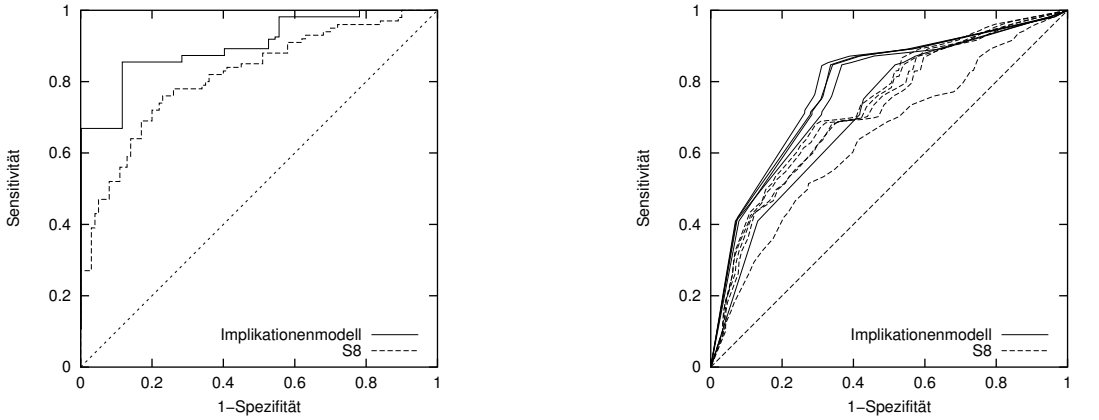
zahl von 5.000 Testdaten erlaubt sind, wurden aus den Originaldaten der Ausgangsbausparkasse fünf verschiedene Testmengen erzeugt. Zur Überprüfung der Generalisierungsfähigkeit wurden aus dem realen Kollektiv der Validierungsbausparkasse weitere fünf Testmengen mit jeweils 5.000 Bauspardarlehen erstellt. Vergleichskriterien waren wiederum die Größen Sensitivität, Spezifität und die AUC–Werte.

7.1.2.2 Klassifikationsergebnisse der Ausgangsbausparkasse

Zur Auswertung des Modells von Truemper wurde das aus Stichprobe S8 erhaltene Regelwerk verwendet, welches die höchsten Klassifikationsergebnisse aufwies. Die

[1] Die Werte beziehen sich auf das Netz N1 mit der höchsten Klassifikationsgüte.

Ergebnisse der weiteren Stichproben waren allerdings nur geringfügig schlechter. Abbildung 7.2 stellt die Ergebnisse einer Testmenge graphisch dar. Die Ergebnisse der vier weiteren Testmengen zeigten einen ähnlichen Verlauf, daher wird auf deren Darstellung verzichtet. Tabelle 7.4 zeigt die zugehörigen Flächeninhalte unter den ROC–Graphen.



(a) Ausgangsbausparkasse

(b) Validierungsbausparkasse

Abbildung 7.2: Vergleich der Klassifikationsergebnisse des modifizierten Modells von Truemper und dem verbandstheoretischen Implikationenmodell der Ausgangs- und Validierungsbausparkasse

Modell	Ausgangsbausparkasse	Validierungsbausparkasse
Modifiziertes Modell von Truemper[2]	0.8035	0.7425
Verbandstheoretisches Implikationenmodell	0.8425	0.7903

Tabelle 7.4: AUC–Werte des verbandstheoretischen Implikationenmodells und des modifizierten Modells von Truemper im Vergleich

7.1.2.3 Klassifikationsergebnisse der Validierungsbausparkasse

Zur Prüfung der Übertragbarkeit wurde ebenfalls das Regelwerk aus Stichprobe S8 verwendet. Die fünf Testmengen wurden mit dem modifizierten Modell von Truemper

[2]Die Werte beziehen sich auf das Regelwerk aus Stichprobe S8, dass die höchste Klassifikationsgüte aufwies.

und dem verbandstheoretischen Implikationenmodell klassifiziert. Die Ergebnisse aller fünf Testmengen sind in Abbildung 7.2 dargestellt. Dabei zeigt die Abbildung, dass bei allen Testmengen das Klassifikationsniveau des modifizierten Modells von Truemper unterhalb des Niveaus des verbandstheoretischen Implikationenmodells liegt.

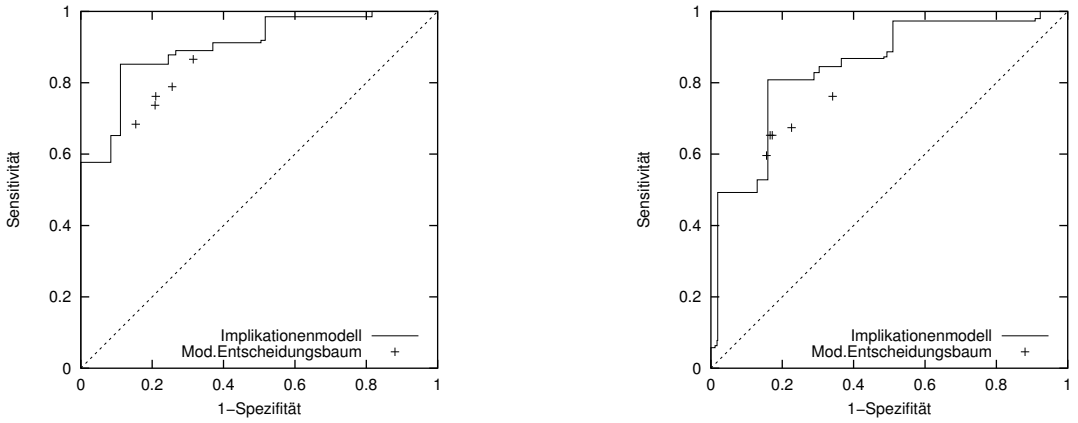
7.1.3 Entscheidungsbäume

7.1.3.1 Datenvorbereitung

Die Eingangsdaten für die Entscheidungsbäume werden ebenfalls mittels der inhaltlichen Diskretisierung transformiert. Analog zum Modell von Truemper umschreiben wir diesen Sachverhalt mit der Bezeichnung modifizierte Entscheidungsbäume. Zur Modellierung verwenden wir die aus den Stichproben S1–S5 erzeugten Entscheidungsbäume, welche die höchste Klassifikationsgüte aufwiesen. Dabei handelte es sich um Stichproben mit hohem Darlehensumfang und einem Mischungsverhältnis von 1:1.

7.1.3.2 Klassifikationsergebnisse der Ausgangsbausparkasse

Abbildung 7.3 zeigt die Klassifikationsergebnisse der modifizierten Entscheidungsbäume der Ausgangsbausparkasse, die nur knapp unter denen des verbandstheoretischen Implikationenmodells liegen.



(a) Ausgangsbausparkasse

(b) Validierungsbausparkasse

Abbildung 7.3: Vergleich der Klassifikationsergebnisse der modifizierten Entscheidungsbäume und dem verbandstheoretischen Implikationenmodell der Ausgangs- und Validierungsbausparkasse

Da es sich bei den Entscheidungsbäumen um ein diskretes Klassifikationsmodell handelt, können keine AUC–Werte angegeben werden. Allerdings sind in [Faw03]

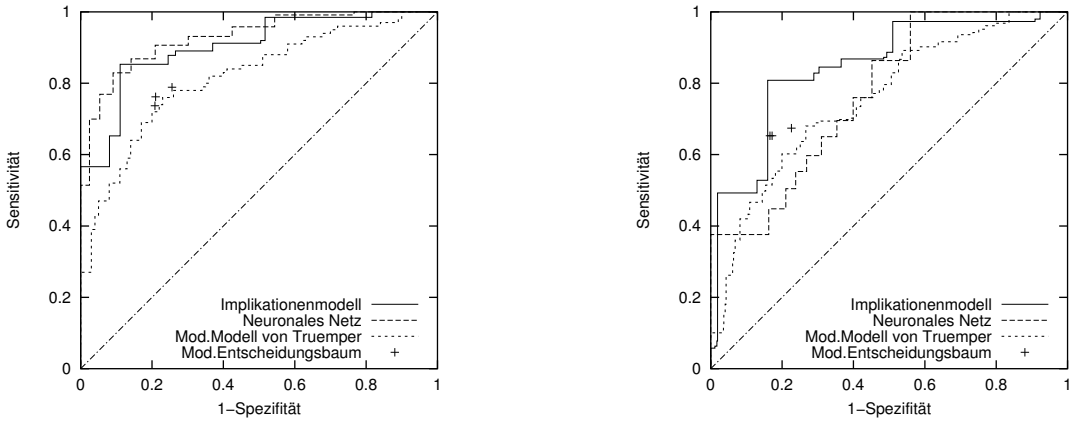
Verfahren beschrieben, um AUC–Werte aus Entscheidungsbäumen zu ermitteln. Für unsere Analysen ist die diskrete Darstellung der Ergebnisse ausreichend, da wir an einer quantitativen Einschätzung der Ergebnisse interessiert sind.

7.1.3.3 Klassifikationsergebnisse der Validierungsbausparkasse

Wie im modifizierten Modell von Truemper verschlechtern sich auch bei den modifizierten Entscheidungsbäumen die Klassifikationsergebnisse bei der Übertragung auf eine weitere Bausparkasse. Abbildung 7.3 zeigt die Klassifikationsergebnisse für die Validierungsbausparkasse. Allerdings fallen die Verschlechterungen in den modifizierten Entscheidungsbäumen deutlich geringer als z. B. in den neuronalen Netzen aus.

7.1.4 Quantitativer Gesamtvergleich aller Klassifikationsmodelle

Im folgenden Abschnitt soll eine abschließende Übersicht der Klassifikationsergebnisse aller vier Modelle dargestellt werden. Dabei sind in Abbildung 7.4 die Ergebnisse aller Modelle für die Ausgangsbausparkasse dargestellt. Aus Gründen der Übersichtlichkeit wurden jeweils die Ergebnisse der Modelle mit der höchsten Klassifikationsgüte ausgewählt. Zudem zeigt Abbildung 7.4 die Ergebnisse der Übertragung aller Modelle auf die Validierungsbausparkasse. Damit kann in allen Modellen die Generalisierungsfähigkeit überprüft werden.



(a) Ausgangsbausparkasse

(b) Validierungsbausparkasse

Abbildung 7.4: Vergleich der Klassifikationsergebnisse aller Modelle der Ausgangs- und Validierungsbausparkasse

Abbildung 7.4 zeigt deutlich, dass das neuronale Netz die höchste Klassifikationsgüte für die Ausgangsbausparkasse liefert, die ROC–Kurve des neuronalen Netzes erreicht das höchste Niveau. Danach folgen das verbandstheoretische Implikationenmodell, die modifizierten Entscheidungsbäume und das modifizierte Modell von Truemper. In drei der vier untersuchten Modelle war eine Diskretisierung der Eingangsdaten nötig; die höchsten Ergebnisse lieferte jedoch das neuronale Netz ohne vorherige Bearbeitung der Eingangsdaten.
Die Betrachtung der Ergebnisse für die Validierungsbausparkasse liefert ein anderes Bild. Hier fallen die neuronalen Netze gerade durch ihr schlechtes Abschneiden auf. Die höchsten Klassifikationsergebnisse liefert das verbandstheoretische Implikationenmodell, gefolgt von den Entscheidungsbäumen und dem Modell von Truemper.

7.1.5 Quantitative Veränderungen in den Modellen

In den folgenden Ausführungen werden die quantitativen Unterschiede in den vier Modellen dargestellt. Dabei werden zuerst die Ergebnisse der Ausgangsbausparkasse analysiert und anschließend die prozentuale Verschlechterung aller Modelle bei der Übertragung auf die Validierungsbausparkasse ermittelt.

- **Ergebnisse der Ausgangsbausparkasse**
  Das neuronale Netz lieferte die höchsten Klassifikationsergebnisse und stellt daher die Basis für unsere Analysen dar. Für die übrigen drei Modelle wird die prozentuale Veränderung der Größen Sensitivität und Spezifität bezogen auf das neuronale Netz ermittelt. Die Abweichungen sind in Tabelle 7.5 dargestellt. Das

	Sensitivität (in %)	Spezifität (in %)
Implikationenmodell	+2.87	-12.11
Modell von Truemper	-8.35	-13.83
Entscheidungsbaum	-8.10	-8.01

Tabelle 7.5: Prozentuale Veränderungen in den Modellen, bezogen auf die Klassifikationsergebnisse des neuronalen Netzes N1

  verbandstheoretische Implikationenmodell weist bei der Sensitivität, also dem Erkennen der ausgefallenen Darlehen, einen höheren Wert als das neuronale Netz auf. Allerdings verschlechtert sich das Implikationenmodell um ca. 12 % beim Erkennen der endgetilgten Bauspardarlehen. Die modifizierten Entscheidungsbäume weisen bei Sensitivität und Spezifität eine gleichbleibende Verschlechterung von ca. 8 % auf. Das modifizierte Modell von Truemper weist ähnliche Veränderungen wie die modifizierten Entscheidungsbäume auf.

- **Übertragung der Modelle auf die Validierungsbausparkasse**
  Im zweiten Schritt soll für jedes Modell die prozentuale Veränderung bei einer

Übertragung auf die Validierungsbausparkasse ermittelt werden, d. h. jedes Modell wird auf seine Generalisierungsfähigkeiten hin überprüft.
Die Abweichungen in Tabelle 7.6 zeigen, dass das verbandstheoretische Implikationenmodell bei Sensitivität und Spezifität konstante, relativ geringe Verschlechterungen von ca. 5 % aufweist. Die Veränderungen der drei anderen Modellen fallen unterschiedlich aus. Das modifizierte Modell von Truemper und die modifizierten Entscheidungsbäume zeigten sogar eine Verbesserung beim Erkennen der endgetilgten Bausparkonten. Demgegenüber steht jedoch beim modifizierten Modell von Truemper eine Verschlechterung beim Erkennen von ausgefallenen Darlehen von nahezu 30 %. Die Abweichung fällt bei den modifizierten Entscheidungsbäumen etwas geringer aus. Die größten Abweichungen bei der Generalisierung liefert das neuronale Netz.

	Sensitivität (in %)	Spezifität (in %)
Implikationenmodell	-5.24	-5.76
Neuronales Netz	-33.36	-14.84
Modell von Truemper	-27.14	+8.15
Entscheidungsbaum	-14.30	+4.81

Tabelle 7.6: Veränderungen in den Klassifikationsergebnissen bei der Übertragung des jeweiligen Modells auf die Validierungsbausparkasse

Das verbandstheoretische Implikationenmodell liefert also robuste Klassifikationsergebnisse für die Ausgangs- und Validierungsbausparkasse.

7.2 Qualitativer Vergleich

Im folgenden Abschnitt wird näher auf die Ursachen der mangelnden Übertragbarkeit der drei Modelle im Vergleich zum verbandstheoretischen Implikationenmodell eingegangen. Die Abweichungen in den Klassifikationsergebnissen werden zudem inhaltlich analysiert. Daneben werden inhaltliche Gemeinsamkeiten und Unterschiede der Modelle untersucht.

7.2.1 Neuronale Netze

7.2.1.1 Interpretation der Abweichungen

Auffallend bei der Modellierung mit neuronalen Netzen war die bessere Klassifikationsgüte für die Ausgangsbausparkasse und die sehr schlechte Übertragbarkeit des Modells auf die Validierungsbausparkasse. Zur Begründung dieses Verhaltens werden wir vor allem auf die Gewichte der neuronalen Netze bei den einzelnen Merkmalen

eingehen und das Auftreten der entsprechenden Merkmale bei beiden Bausparkassen untersuchen.

Tabelle 7.7 zeigt ausgewählte Gewichte der neuronalen Netze 1–4. Da die Höhe der Gewichte maßgeblich für die Klassifikation eines Datensatzes ist, wurden nur Merkmale berücksichtigt, deren Gewichte α_i zumindest in einem verdeckten Knoten y_i größer als 2, bzw. kleiner als -2 waren. Die Kennzeichnung * bedeutet, dass kein Gewicht α_i zum verdeckten Knoten y_i vorhanden ist oder der Einfluss des Knotens zu gering ist.

	Netz 1				**Netz 2**			**Netz 3**				**Netz 4**		
	y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_1	y_2	y_3	y_4	y_1	y_2	y_3
WK[3]	*	*	*	*	-6.3	*	*	*	-0.9	-2.5	3.9	-3.1	-0.4	*
WS	*	*	7.1	*	9	-6.1	6.7	*	12.5	*	*	*	*	10.8
DH	-0.1	*	*	-5.3	*	*	-3.5	-1.9	-0.6	-2.3	*	-2.5	-0.3	*
T1	-3.4	-3.4	*	*	*	-9.5	-5.2	6.8	1.1	-9.7	-4.5	-1.2	-4.1	1.1
T2	-3.6	-2.5	*	*	*	-7.6	*	*	*	*	8	*	-4.3	*
T3	*	*	-3.7	*	7.8	3.6	*	*	*	*	*	4.6	*	*
T4	*	*	-4.8	*	*	*	*	*	*	*	*	*	*	*
AB	*	*	*	*	*	*	*	*	*	*	*	2.3	*	*
BE	-3.5	*	*	*	*	*	*	*	-2.8	*	*	*	*	*
SP	-8.4	*	11.4	*	*	-24.5	-30.8	17.1	*	-38.5	*	*	-11.1	-12.5
WP	*	8.1	*	*	*	*	*	*	-10.9	4.3	*	*	*	2.1
AL	*	*	*	-4.8	*	*	-4.8	*	*	*	8.8	*	*	*

Tabelle 7.7: Ausgewählte Gewichte der vier neuronalen Netze

Dabei ist auffallend, dass die Berufsgruppen Angestellter, Selbständig, Rentner und sonstige Berufe im neuronalen Netz einen äußerst geringen Einfluss besitzen. Im verbandstheoretischen Implikationenmodell werden sie hingegen zur Klassifikation verwendet. Da hierin eine mögliche Ursache für die schlechte Übertragbarkeit des neuronalen Netzes liegen könnte, wird das Auftreten dieser Berufsgruppen in beiden Bausparkassen untersucht. Weiterhin kann man beobachten, dass die Berufsgruppe Arbeiter in nur einem Netz mit relativ geringem Gewicht auftritt. Demgegenüber findet sich die Berufsgruppe Arbeiter verstärkt im verbandstheoretischen Implikationenmodell zur Erkennung ausgefallener Darlehen. Auch die Berufsgruppe Beamte wird, im Gegensatz zum verbandstheoretischen Implikationenmodell, in nur zwei von vier Netzen zur Klassifikation verwendet. Diese Beobachtungen liefern einen ersten Ansatzpunkt zur Analyse. Schwierig ist es jedoch, Aussagen über rationale Variablen wie Darlehenshöhe, Spardauer, Wohnungsbauprämie oder Alter zu treffen, da diese nicht diskretisiert werden müssen. Daher beschränken wir unsere Analysen auf binäre Merkmale.

Wir wollen in den folgenden Ausführungen die schlechte Übertragbarkeit der neu-

[3]Dabei bezeichnen die Kürzel: WK=Weiteres Darlehenskonto nicht in Zahlungsschwierigkeiten, WS=Weiteres Darlehenskonto in Zahlungsschwierigkeiten, DH=maximal möglicher Darlehensanspruch, T1–4=Tarife 1–4, AB=Arbeiter, BE=Beamte, SP= Spardauer, WP=WoP und AL=Alter.

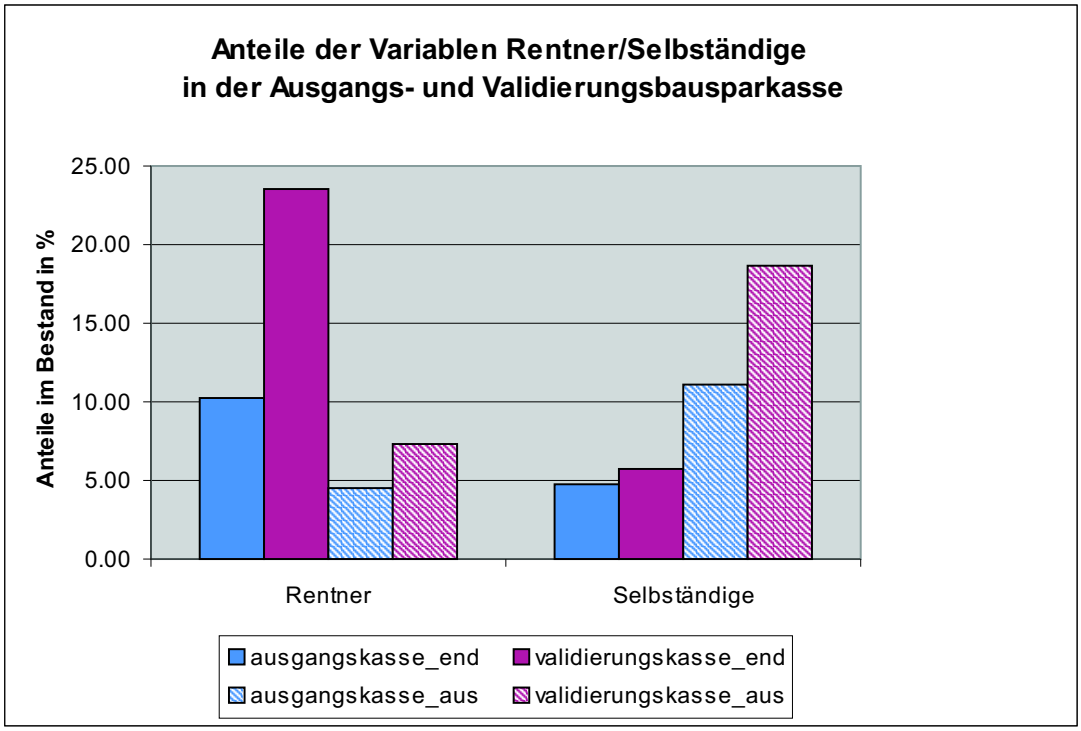


Abbildung 7.5: Anteile der Berufsgruppen Rentner und Selbständige im Bestand der Ausgangs- und Validierungsbausparkasse

ronalen Netze mit Hilfe der verwendeten Variablen in den Modellen analysieren. Die unterschiedliche Verwendung, sowie die unterschiedlichen Häufigkeiten des Auftretens dieser Merkmale, könnten für die stark schwankenden Klassifikationsergebnisse verantwortlich sein. In den neuronalen Netzen müssen Variablen vorhanden sein, deren Gewichte null, bzw. äußerst gering sind. Diese Variablen sollten im verbandstheoretischen Implikationenmodell verwendet werden, zudem sollte sich die Häufigkeit des Auftretens in der Validierungsbausparkasse stark von der Häufigkeit in der Ausgangsbausparkasse unterscheiden. Dies trifft auf die Variablen Rentner und Selbständige zu, wie Abbildung 7.5 zeigt. Dabei bezeichnet „ausgangskasse_end" und „validierungskasse_end" die endgetilgten Bauspardarlehen der Ausgangs- und Validierungsbausparkasse. Analoges gilt für die Bezeichnungen „ausgangskasse_aus" und „validierungskasse_aus".
Das Merkmal Rentner ist in den endgetilgten Darlehen der Validierungsbausparkasse mehr als doppelt so häufig vorhanden wie in der Ausgangsbausparkasse. Im verbandstheoretischen Implikationenmodell wird das Merkmal zur Erkennung endgetilgter Bauspardarlehen verwendet. Ein ähnlicher Zusammenhang ist beim Merkmal Selbständige zu entdecken, das im neuronalen Netz nahezu keinen Einfluss hat. In den

ausgefallenen Darlehen der Validierungsbausparkasse ist es nahezu doppelt so häufig anzutreffen als in der Ausgangsbausparkasse. Dieses Merkmal dient im verbandstheoretischen Implikationenmodell zum Klassifizieren ausgefallener Darlehen. Die Analysen bieten daher einen Anhaltspunkt zur Interpretation der Abweichungen bei der Übertragung des neuronalen Netzes auf die Validierungsbausparkasse.

7.2.1.2 Zusammenfassung

Wir wollen im folgenden Abschnitt die Vor- und Nachteile bei der Verwendung von neuronalen Netzen zur Klassifikation von Bauspardarlehen eingehen.
Die neuronalen Netze lieferten für die Ausgangsbausparkasse die höchste Klassifikationgüte, waren also den logischen Modellen überlegen. Zudem entfällt bei der Verwendung von neuronalen Netzen die oftmals schwierige Diskretisierung von rationalen Merkmalen. Ein weiterer Vorteil der neuronalen Netze liegt darin, dass sie in der Lage sind, alle Darlehen zu klassifizieren. Das verbandstheoretische Implikationenmodell gewährleistet hingegen keine vollständige Bewertung aller Darlehen.
Allerdings liegt das größter Manko der neuronalen Netze in der mangelnden Generalisierungsfähigkeit der Ergebnisse. Die logischen Modelle wiesen deutlich geringe Abweichungen bei der Übertragung auf die Validierungsbausparkasse auf. Ein weiterer Nachteil ist die schlechte Interpretierbarkeit der neuronalen Netze. Das Klassifikationsergebnis ist für den Anwender häufig nicht nachvollziehbar. Hingegen versuchen die logischen Modelle interpretierbare Regeln aus den Trainingsbeispielen zu erzeugen.

7.2.2 Modell von Truemper

7.2.2.1 Interpretation der Abweichungen

Zur Ermittlung einer trennenden DNF–Formel werden im Modell von Truemper alle Merkmale verwendet. Demgegenüber werden im verbandstheoretischen Implikationenmodell von Beginn an signifikante Strukturen für die Klassifikation herausgearbeitet. Daher ist es möglich, dass die trennende DNF–Formel Variablen enthält, die im verbandstheoretischen Implikationenmodell nicht zur Bewertung verwendet werden. Wir wollen in den folgenden Analysen diese Merkmale hinsichtlich ihrer relativen Häufigkeit in der Ausgangs- und Validierungsbausparkasse untersuchen, da sie für die Abweichungen bei der Übertragung im Modell von Truemper verantwortlich sein können.
Bei Betrachtung des verbandstheoretischen Implikationenregelwerkes fällt auf, dass die Merkmale mittleres und hohes Darlehen, Tarif 2 und 4, Beruf Angestellter und mittlere und hohe Wohnungsbauprämie gar nicht, oder äußerst selten auftreten. Daher wurden die relativen Häufigkeiten des Auftretens dieser Merkmale in den Testmengen der Ausgangs- und Validierungsbausparkasse untersucht. Tabelle 7.6 zeigt, dass das

modifizierte Modell von Truemper eine deutlich schlechtere Sensitivität bei der Generalisierung aufweist, die Anzahl der vom Modell erkannten ausgefallenen Darlehen ist also geringer. Daher beschränken wir uns in der Analyse auf ausgefallene Bauspardarlehen. Abbildung 7.6 zeigt die Ergebnisse dieser Untersuchung. Die Merkmale

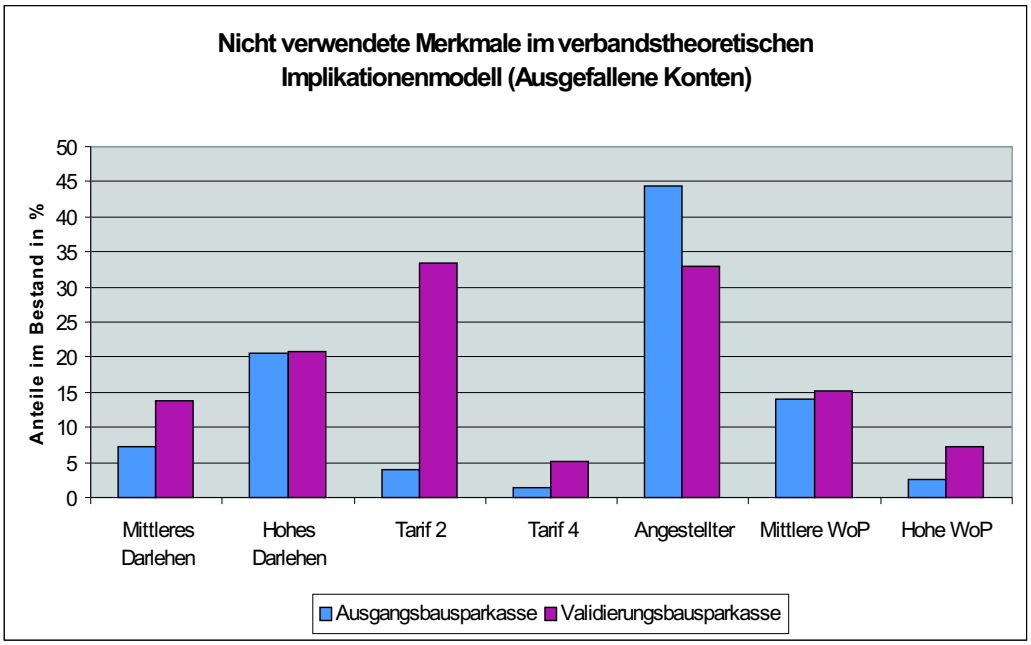


Abbildung 7.6: Verteilung der nicht verwendeten Merkmale bei den ausgefallenen Bausparkonten im verbandstheoretischen Implikationenmodell

hohes Darlehen und mittlere Wohnungsbauprämie kommen bei beiden Bausparkassen annähernd gleich häufig vor und können daher nicht für das schlechtere Klassifikationsergebnis verantwortlich sein. Die anderen Merkmale weisen jedoch große Unterschiede auf. So ist der Bezug einer hohen Wohnungsbauprämie in der Validierungsbausparkasse nahezu dreimal häufiger anzutreffen als in der Ausgangsbausparkasse. In der Validierungsbausparkasse sind doppelt so viele Konten mit mittlerer Darlehenssumme vorhanden. Die Tarife 2 und 4 sind in der Validierungsbausparkasse bei den ausgefallenen Bauspardarlehen neunmal, bzw. dreimal häufiger anzutreffen als in der Ausgangsbausparkasse. Einzig die Berufsgruppe Angestellte ist seltener in der Validierungsbausparkasse vorhanden.
Alle genannten Merkmale werden im Regelwerk des verbandstheoretischen Implikationenmodells nicht verwendet, da sie sich bei den Signifikanzprüfungen als nicht relevant erwiesen haben. Im modifizierten Modell von Truemper finden sie jedoch Eingang und tragen damit zur Bewertung des Darlehens bei. Sie sind daher möglicherweise auch für die Verschlechterung der Sensitivität verantwortlich.

7.2.2.2 Gemeinsamkeiten mit dem verbandstheoretischen Implikationenmodell

Das Modell von Truemper weist trotzdem inhaltliche Gemeinsamkeiten zum verbandstheoretischen Implikationenmodell auf. Diese sollen im folgenden kurz erläutert werden.

- **Suchen und Bewerten von logischen Regeln**
  Beide Modelle suchen nach logischen Regeln in den Daten, die eine Klassifikation ermöglichen. Damit unterscheiden sich die beiden Modelle fundamental von den probabilistischen Ansätzen, wie z. B. den neuronalen Netzen.

- **Bereitstellung der Eingangsdaten**
  Beide Modelle verlangen eine Diskretisierung der rationalen Eingangsdaten. Im Modell von Truemper wird die „Cutpoint–Methode" verwendet, während die Diskretisierung im verbandstheoretischen Implikationenmodell mit Hilfe von Expertenwissen stattfindet.

- **Regelstruktur**
  Die Anzahl der Regeln im verbandstheoretischen Implikationenmodell ist zwar deutlich geringer als im Modell von Truemper, dennoch finden sich Regeln in ähnlicher Form im modifizierten Modell von Truemper wieder. Beispiele dafür sind in Tabelle 7.8 dargestellt.

7.2.2.3 Unterschiede zum verbandstheoretischen Implikationenmodell

Trotz der Gemeinsamkeiten bestehen zwischen den beiden Modellen inhaltliche Unterschiede, wie die folgenden Ausführungen zeigen.

- **Anzahl der Regeln und Interpretierbarkeit**
  Die Anzahl der erzeugten Formeln im Modell von Truemper ist sehr hoch. Dadurch leidet vor allem die Interpretierbarkeit des Modells. Für den Anwender ist nicht nachvollziehbar, welche Formel letztendlich den Ausschlag für das Klassifikationsergebnis gegeben hat.

- **Art der Bewertung der Regeln**
  Im Modell von Truemper werden die erfüllten Formeln nicht nach ihrer Stärke, sondern allein nach ihrer Anzahl bewertet. Jede erfüllte Formel entspricht damit genau einer Stimme, die entsprechend für die Zugehörigkeit zur Klasse A oder B gewertet wird. Im Gegensatz dazu werden im verbandstheoretischen Implikationenmodell die Regeln mit ihrer individuellen Stärke für das Klassifikationsergebnis bewertet.

- **Art der Untersuchung**
  Bei Formulierung der „Clash– und Agree–Bedingungen" im Modell von Truemper werden alle Variablen verwendet. Hingegen versucht das verbandstheoretische Implikationenmodell die innere Struktur bzw. den inneren Aufbau in den Daten zu finden. Dies führt dazu, dass für die Klassifikation nicht relevante Merkmale herausgefiltert werden. Das Herausarbeiten wichtiger Merkmalskombinationen ist jedoch Bestandteil unserer Zielformulierung.

Verbandstheoretisches Implikationenmodell	Modifiziertes Modell von Truemper
Ausgefallene Darlehen	
WK–S	WK–S
Selbständig → Tarif 1/Tarif 3	Selbständig ∧ Tarif 2
Selbständig ∧ Hohes Darlehen → Geringe WoP	Selbständig ∧ Hohes Darlehen ∧ alt
Selbständig ∧ Geringe Spard. → Geringe WoP	Selbständig ∧ Geringe Spard. ∧ Tarif 2
Arbeiter ∧ Mittl. Spard. → Tarif 1/Tarif 3	Arbeiter ∧ Mittl. Spard. ∧ Mittl. Alter Arbeiter ∧ Hohes Darlehen ∧ Mittl. Spard.
Hohe Spard. ∧ Mittl. Alter → Geringe WoP	Hohe Spard. ∧ Mittl. Alter ∧ Geringe WoP
Endgetilgte Darlehen	
WK–G	WK–G ∧ Tarif 2
WK–G	WK–G ∧ Hohes Darlehen ∧ alt
Geringe Spard. ∧ alt → Kleines Darlehen	Geringe Spard. ∧ alt
Geringe Spard. ∧ alt ∧ K–Angest → K–Beamter	Geringe Spard. ∧ alt ∧ K–Angest
Kleines Darlehen ∧ N–Selbst ∧ alt → K–Beamter	Kleines Darlehen ∧ K–Arbeiter ∧ alt ∧ N–Selbst ∧ Mittl. Spard. ∧ K–Rentner
Beamter → Tarif 1/Tarif 3	Beamter ∧ Mittl. Spard.
Rentner → Tarif 1/Tarif 3	Rentner ∧ Mittl.WoP

Tabelle 7.8: Ähnliche Regeln für ausgefallene und endgetilgte Darlehen im verbandstheoretischen Implikationenmodell sowie im modifizierten Modell von Truemper

7.2.3 Entscheidungsbäume

7.2.3.1 Interpretation der Abweichungen

Die Entscheidungsbäume und das verbandstheoretische Implikationenmodell besitzen ausgeprägte inhaltliche Gemeinsamkeiten. Dies dürfte der Grund sein, dass die Entscheidungsbäume relativ robuste Klassifikationsergebnisse für die Ausgangs- und Validierungsbausparkasse liefern. Im folgenden Abschnitt soll die Frage erläutert wer-

den, warum die Entscheidungsbäume dennoch bei der Übertragung auf die Validierungsbausparkasse schlechtere Ergebnisse beim Erkennen der ausgefallenen Darlehen erzielen. Entsprechend der Argumentation bei den neuronalen Netzen und dem modifizierten Modell von Truemper werden Merkmale betrachtet, die im verbandstheoretischen Implikationenmodell verwendet werden, bzw. dort nicht berücksichtigt sind. Die relative Häufigkeit dieser Merkmale wird in der Ausgangs- und Validierungsbausparkasse untersucht.
Die Analyse von fünf modifizierten Entscheidungsbäumen mit der höchsten Klassifikationsgüte zeigt, dass in keinem der Bäume die Merkmale Rentner und Beamter verwendet wurden. Weiterhin wird die Variable Darlehenshöhe mit ihren Ausprägungen gering, mittel und hoch in nur einem der fünf Bäume zur Klassifikation verwendet. Die Tarife 2 und 4 werden dagegen im verbandstheoretischen Implikationenmodell nicht verwendet. Eine Übersicht der Verwendung dieser Merkmale ist in Tabelle 7.9 dargestellt. Abbildung 7.7 zeigt, dass es durchaus Unterschiede im Auftreten der

Merkmal	mod. Entscheidungsbaum	Implikationenmodell
Rentner	nicht vorhanden	vorhanden
Beamter	nicht vorhanden	vorhanden
Geringes Darlehen	selten vorhanden	vorhanden
Hohes Darlehen	selten vorhanden	vorhanden
Tarif 2	vorhanden	nicht vorhanden
Tarif 4	vorhanden	nicht vorhanden

Tabelle 7.9: Betrachtung von verschiedenen Merkmalen hinsichtlich ihrer Verwendung beim entsprechenden Klassifikationsmodell

Merkmale bei beiden Bausparkassen gibt. Dabei bezeichnet „ausgangskasse_end" und „validierungskasse_end" die endgetilgten Bauspardarlehen der Ausgangs- und Validierungsbausparkasse. Analoges gilt für die Bezeichnungen „ausgangskasse_aus" und „validierungskasse_aus". In den endgetilgten Darlehen der Validierungsbausparkasse ist die Berufsgruppe Rentner mehr als doppelt so häufig vorhanden wie in den endgetilgten Darlehen der Ausgangsbausparkasse. Diese Auffälligkeit ist bei den ausgefallenen Darlehen nicht ganz so stark ausgeprägt. Deutlicher sind die Unterschiede in den Tarifen 2 und 4 zu erkennen. In der Validierungsbausparkasse ist der Anteil der ausgefallenen Darlehen im Tarif 2 neunmal so hoch wie in der Ausgangsbausparkasse. Das Merkmal Tarif 2 wird in drei der fünf Bäumen verwendet um endgetilgte Darlehen zu klassifizieren. Bauspardarlehen, die das Merkmal Tarif 2 besitzen, werden daher häufig als endgetilgt bewertet, obwohl sie eigentlich ausgefallen sind. Dies führt zu einer Verschlechterung der Sensitivität und bietet einen Erklärungsansatz für die schlechtere Generalisierungsfähigkeit.

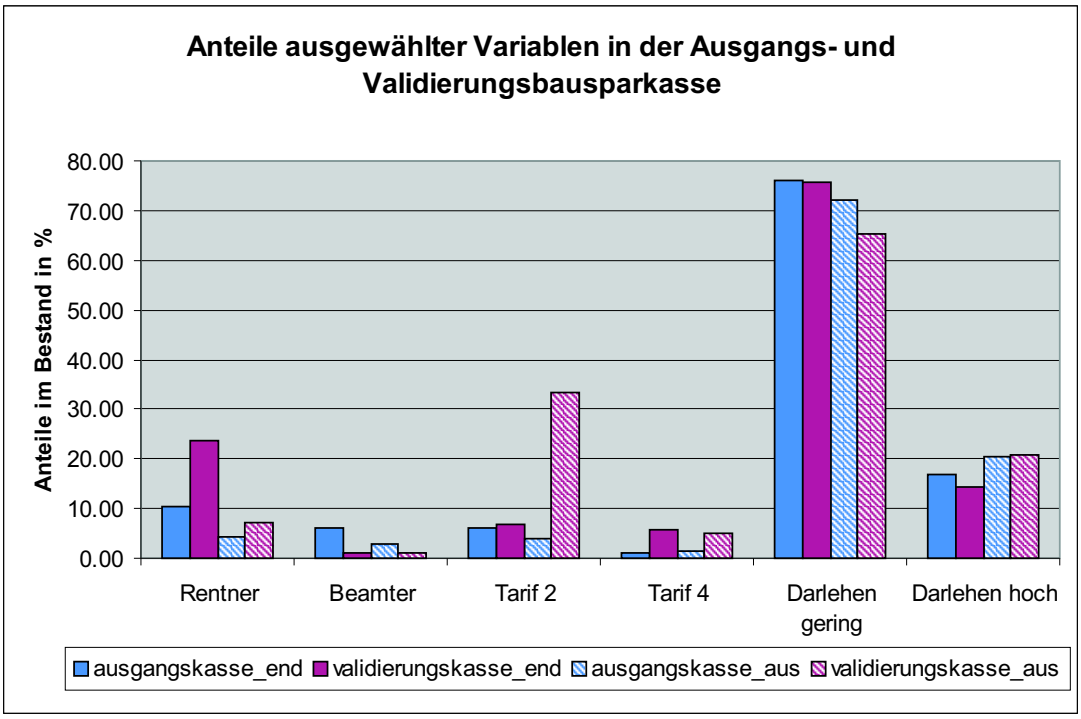


Abbildung 7.7: Relative Häufigkeiten ausgewählter Variablen in der Ausgangs- und Validierungsbausparkasse differenziert nach ausgefallenen und endgetilgten Bauspardarlehen

7.2.3.2 Gemeinsamkeiten mit dem verbandstheoretischen Implikationenmodell

- **Relevanz der Merkmale**
  Entscheidungsbäume beginnen an der Wurzel mit dem Merkmal, dass den höchsten Informationsgewinn liefert, d. h. die Relevanz der Merkmale wird berücksichtigt. Merkmale, die keinen signifikanten Beitrag zur Klassifizierung leisten, werden im Entscheidungsbaum nicht verwendet. Analoges gilt für Merkmale bzw. Regeln im verbandstheoretischen Implikationenmodell.

- **Bereitstellung der Eingangsdaten**
  Beide logischen Modelle verlangen eine Diskretisierung der Eingangsdaten. Dies geschieht bei den Entscheidungsbäumen mit Hilfe eines entropiebasierten Ansatzes.

- **Regelstruktur**
  Die in den modifizierten Entscheidungsbäumen verwendeten Regeln finden sich auch im verbandstheoretischen Implikationenmodell wieder. Der modifizierte

Baum in Abbildung 7.8 enthält Strukturen, die in ähnlicher Form auch im verbandstheoretischen Implikationenmodell zu finden sind. Die Modellierung mit Entscheidungsbäumen liefert also ähnliche Strukturen wie Tabelle 7.10 zeigt.

- **Interpretierbarkeit**
  Jeder Pfad vom Wurzel- bis zu einem Klassifizierungsknoten liefert eine Entscheidungsregel. Dies gewährleistet eine hohe Interpretierbarkeit für den Anwender.

Verbandstheoretisches Implikationenmodell	Modifizierte Entscheidungsbäume
WK–S	WK–S → Risiko hoch
Selbständig → Tarif 1/Tarif 3	Selbständig → Risiko hoch
WK–G	WK–G → Risiko gering
N–Selbst. ∧ Geringe Spard. ∧ alt → K–Beamter	N–Selbst. ∧ Geringe Spard. ∧ alt → Risiko gering

Tabelle 7.10: Ähnliche Regelstrukturen im verbandstheoretischen Implikationenmodell und den modifizierten Entscheidungsbäumen

7.2.3.3 Unterschiede zum verbandstheoretischen Implikationenmodell

- **Kombinationsmöglichkeiten der Regeln**
  Die Regeln eines Pfades können nicht mit Regeln aus anderen Pfaden kombiniert werden, da zur Klassifikation der Entscheidungsbaum vom Wurzelknoten beginnend nach unten abgearbeitet wird. Das Regelwerk im verbandstheoretischen Implikationenmodell ist kombinierbar und wird auch entsprechend des kombiniertem Auftretens bewertet.

- **Interpretierbarkeit**
  Die Regeln werden bei zunehmender Größe des Baumes komplexer. Dies kann dazu führen, dass einzelne Regeln, abhängig von der Tiefe des Baumes, mehr als zehn Merkmale umfassen. Damit leidet jedoch die Interpretierbarkeit der Entscheidungsbäume.

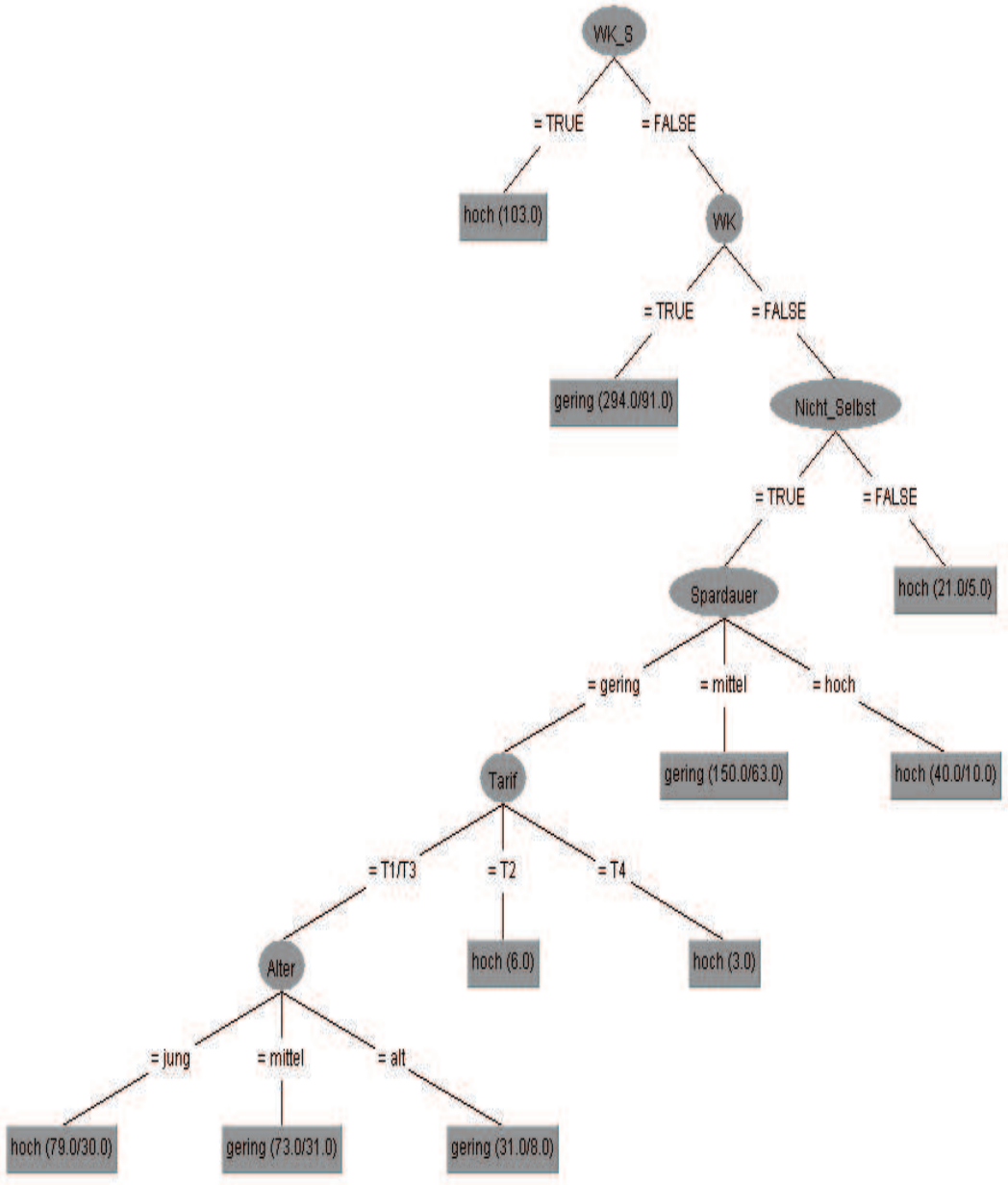


Abbildung 7.8: Entscheidungsbaum einer nach dem verbandstheoretischen Implikationenmodell diskretisierten Stichprobe von 800 Konten

Kapitel 8

Weitere Anwendungsmöglichkeiten des verbandstheoretischen Implikationenmodells

8.1 Klassifizierung von Darlehensverzichtern

8.1.1 Untersuchungsaufbau

Nach Beendigung der Sparphase hat der Bausparer Anspruch auf sein Bauspardarlehen [TCDL02]. Die Inanspruchnahme des Darlehens hängt allerdings von vielen Faktoren ab. Ein entscheidender Faktor ist der Außenzins, wie z. B. in [Che05] gezeigt wurde.
Die Ermittlung weiterer Faktoren für den Darlehensverzicht bzw. die Darlehensnahme sollen hier im Vordergrund stehen. Die Entscheidung des Bausparers könnte beispielsweise von seinem Beruf oder der Spardauer beeinflusst werden. Eventuell deuten bestimmte Merkmalskombinationen auf einen Darlehensverzicht hin. Um die zusätzlichen Faktoren zu ermitteln, sollte allerdings der Zinseinfluss ausgeblendet werden. Dies wird dadurch erreicht, indem ausschließlich Datensätze eines Jahres mit geringen Zinsschwankungen gewählt wurden. Als Vergleichszins dient der 10–Jahres–Hypothekenzins[1] der Deutschen Bundesbank, der im Jahr 2000 nur geringe Schwankungen aufwies, wie Tabelle 8.1 zeigt. Die Validierungsbausparkasse wurde als Referenzbausparkasse für die Untersuchung gewählt.
Es wurden in diesem Zeitraum nur Konten betrachtet, die keinen zeitlichen Verschub bei Auszahlung des Guthabens und des Darlehens aufweisen. Ein Auszahlungsverschub kann darauf hindeuten, dass der Darlehensnehmer seine Entscheidung über die Inanspruchnahme des Bauspardarlehens vom aktuellen Zinsniveau am Kapitalmarkt

[1] Quelle: `http://www.bundesbank.de/statistik/statistik_zeitreihen.php?func=row&tr=su0046`

abhängig macht. Er rechnet mit fallenden Zinsen und verzögert daher die Auszahlung des Darlehens. Aus diesem Grund müssen Monat und Jahr der Zuteilung, Guthabens- und Darlehensauszahlung bzw. des Darlehensverzichts bei den ausgewählten Bausparkonten identisch sein. Da die vier Tarife der Validierungsbausparkasse unterschiedliche Darlehenszinsen aufweisen, wurden zur Untersuchung ausschließlich reine Bauspardarlehen (ohne VK/ZK–Konten) des Tarifes 1 ausgewählt. Das Auswahlverfahren

Zeitpunkt (Jahr-Monat)	Wert (% p.a.)
2000-12	6.44
2000-11	6.64
2000-10	6.68
2000-09	6.72
2000-08	6.69
2000-07	6.70
2000-06	6.64
2000-05	6.73
2000-04	6.54
2000-03	6.64
2000-02	6.76
2000-01	6.69

Tabelle 8.1: Sollzinsen Banken/Hypothekarkredite auf Wohngrundstücke zu Festzinsen auf 10 Jahre (Effektivzinssatz, Durchschnittszinssatz)

lieferte 4.483 Darlehensverzichter und 4.673 Darlehensnehmer im Bestand der Validierungsbausparkasse für das Jahr 2000. Der Gesamtbestand umfasst 75.827 Darlehensverzichter und 128.152 Darlehensnehmer der Jahre 1991–2005 im Tarif 1. In Anlehnung an die Modellierung der Kreditausfallwahrscheinlichkeiten wurden folgende Merkmale zur Analyse verwendet:

- Spardauer in Jahren
- Weiterer Bausparvertrag vorhanden
- WoP–Bezug
- WoP–Höhe
- Beruf des Bausparers
- Alter des Bausparers
- Bausparsumme
- VL–Bezug

Für die Merkmale Spardauer, WoP–Höhe, Alter und Bausparsumme wurde die inhaltliche Diskretisierung aus Unterabschnitt 5.6.2 gewählt. Die Merkmale VL–Bezug, WoP–Bezug, Berufsgruppe und weiterer Bausparvertrag vorhanden wurden binär kodiert. Die Daten wurden in eine Trainingsmenge zur Erstellung der logischen Regeln und eine Testmenge unterteilt. Die Trainingsmenge wurde aufgeteilt in die Menge D^+, die alle Konten enthält, die das Zielattribut Darlehensverzicht aufwiesen, und eine Menge D^-, die alle Darlehen enthält die das Zielattribut nicht besitzen, also ihr Darlehen in Anspruch genommen haben.

Das verbandstheoretische Implikationenmodell soll in der Lage sein, mit Hilfe von signifikanten Implikationen die Darlehensverzichter und Darlehensnehmer zu klassifizieren. Weiterhin sollen Darlehensverzichterwahrscheinlichkeiten ermittelt werden, die es erlauben, ungesehene Bausparverträge hinsichtlich ihrer Entscheidung über Darlehensannahme oder Darlehensverzicht zu bewerten.

8.1.2 Ermittlung der Stammbasis und signifikanter Implikationen

Aus den Mengen D^+ und D^- wurden zwei formale Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ erzeugt und mit Hilfe des Programms ConImp die Stammbasis der Implikationen für die Kontexte ermittelt. Um eine Reduzierung der Implikationen in der Stammbasis zu erreichen, wurden zum einen nur Implikationen mit erfüllter Prämisse betrachtet, zum anderen wurde der Schwellwert $\delta \in \mathbb{R}$ analog zur Vorgehensweise bei der Klassifizierung ausgefallener Darlehen auf 10 % gesetzt. Es werden also nur Implikationen zur weiteren Analyse verwendet, die von mindestens 10 % der Bausparkonten echt erfüllt wurden.

Allerdings stellt sich die Frage, inwieweit die so ermittelten Implikationen auch signifikant für die Mengen D^+ und D^- sind. Eventuell sind Implikationen in beiden Mengen gleichmäßig vorhanden. Um die Signifikanz und damit die Relevanz der Implikationen für Darlehensverzicht und Darlehensnahme zu untersuchen, wird der χ^2–Unabhängigkeitstest aus Unterabschnitt 4.5 durchgeführt. Die Nullhypothese H_0 (die Merkmale sind stochastisch unabhängig) wird mit einem Signifikanzniveau $\alpha = 1\,\%$ abgelehnt, wenn für die Prüfgröße T gilt:

$$T = n \cdot \frac{(N_{11} \cdot N_{22} - N_{12} \cdot N_{21})^2}{N_{1\bullet} \cdot N_{2\bullet} \cdot N_{\bullet 1} \cdot N_{\bullet 2}} > \chi^2_{1;0.99}$$

Das zugehörige (1 - α)–Quantil der χ^2–Verteilung beträgt 6.64 wie z. B. [BLK06] zu entnehmen ist. Für eine Implikation wird beispielhaft der χ^2–Wert mit Hilfe einer Kontingenztabelle ermittelt.

Beispiel 8.1. *Für die Implikation VL → WK soll überprüft werden, inwieweit sie für die Struktur der Darlehensverzichter signifikant ist. Die empirische Ermittlung lieferte folgende Kontingenztabelle:*

VL → WK	D^+	D^-	$\sum$
vorhanden	445	101	546
nicht vorhanden	1.555	1.899	3.454
$\sum$	2.000	2.000	4.000

Die Prüfgröße T wird daraus folgendermaßen ermittelt:

$$\begin{aligned} T &= 4000 \cdot \frac{(445 \cdot 1899 - 1555 \cdot 101)^2}{2000 \cdot 2000 \cdot 546 \cdot 3454} \\ T &\approx 251 > \chi^2_{1;0.9} = 6.64 \end{aligned}$$

$\Longrightarrow$ *Die Nullhypothese der stochastischen Unabhängigkeit kann damit verworfen werden. Das Merkmal ist stochastisch abhängig, d. h. die Implikation tritt in der Menge D^+ signifikant häufiger auf und kann daher zur Klassifikation der Darlehensverzichter verwendet werden.*

Die Signifikanzprüfung wurde für alle Implikationen der Stammbasis der Kontexte $\mathbb{K}^+$ und $\mathbb{K}^-$ durchgeführt.

8.1.3 Erstellung der Regelsätze

Aus dem vorigen Schritten erhält man zwei Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ von signifikanten Implikationen mit zugehörigen χ^2–Werten für die Struktur der Mengen D^+ und D^-. Allerdings müssen daraus handhabbare Regelsätze entstehen, die in der Lage sind, die Trainingsbeispiele gut zu klassifizieren und dabei eine hohe Anzahl von Datensätzen überdecken.
Die Implikationen der Mengen $\mathcal{L}_1$ und $\mathcal{L}_2$ wurden mit Hilfe des Medians in vier Teilmengen $\mathcal{L}_{11}, \mathcal{L}_{12}, \mathcal{L}_{21}$ und $\mathcal{L}_{22}$ in disjunktiver Normalform zerlegt, für die gilt:

$\mathcal{L}_{1\bullet}$ = Regeln, die aus dem Kontext $\mathbb{K}^+$ ermittelt wurden,
$\mathcal{L}_{2\bullet}$ = Regeln, die aus dem Kontext $\mathbb{K}^-$ ermittelt wurden.

Die Teilmengen umfassen folgende Implikationen:

- $\mathcal{L}_{11}$: Stark signifikante Regeln des Kontextes $\mathbb{K}^+$

  Geringe Spardauer $\wedge$ WoP–Bezug $\rightarrow$ Weiteres Konto vorhanden
  VL–Bezug $\rightarrow$ Weiteres Konto vorhanden
  Spardauer gering $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden
  WoP–Bezug $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden
  BS gering $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden
  BS gering $\wedge$ WoP–Bezug $\rightarrow$ Weiteres Konto vorhanden
  Mittlere WoP $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden

BS gering ∧ Rentner → alt
Rentner → Weiteres Konto vorhanden

- $\mathcal{L}_{12}$: Signifikante Regeln des Kontextes $\mathbb{K}^+$

  Geringe Spardauer ∧ Mittleres Alter → Weiteres Konto vorhanden
  Angestellter ∧ alt → Weiteres Konto vorhanden
  BS mittel ∧ Geringe Spardauer → Weiteres Konto vorhanden
  Geringe Spardauer ∧ Geringe WoP → Weiteres Konto vorhanden
  Arbeiter ∧ Geringe Spardauer → Weiteres Konto vorhanden
  BS mittel ∧ alt → Weiteres Konto vorhanden
  Geringe WoP ∧ alt → Weiteres Konto vorhanden
  BS gering ∧ Geringe WoP → Weiteres Konto vorhanden
  BS hoch ∧ alt → Weiteres Konto vorhanden
  Angestellter ∧ WoP–Bezug → Weiteres Konto vorhanden

- $\mathcal{L}_{21}$: Stark signifikante Regeln des Kontextes $\mathbb{K}^-$

  Mittlere Spardauer → Weiteres Konto vorhanden
  BS mittel ∧ Mittleres Alter → Weiteres Konto vorhanden
  Mittlere Spardauer ∧ Hohe WoP → WoP–Bezug
  BS mittel ∧ Geringe WoP → Weiteres Konto vorhanden

- $\mathcal{L}_{22}$: Signifikante Regeln des Kontextes $\mathbb{K}^-$

  BS mittel ∧ Hohe WoP → WoP–Bezug
  BS mittel ∧ Angestellter → Weiteres Konto vorhanden
  Hohe WoP ∧ Mittleres Alter → WoP–Bezug
  BS mittel ∧ Arbeiter → Weiteres Konto vorhanden

Innerhalb der Teilmengen sind die Implikationen nach ihrem χ^2–Wert absteigend sortiert. Auffallend ist, dass die Struktur der Darlehensverzichter vor allem durch die Merkmale geringe Spardauer, Bezug von vermögenswirksamen Leistungen, hohes Alter und weiterer Bausparvertrag vorhanden geprägt ist. Dies legt folgende Schlüsse nahe:

- **Hohes Alter**
  Der Bausparer scheut aufgrund seines Alters eventuell die Inanspruchnahme des Darlehens, sei es z. B. aus Angst vor Arbeitsplatzverlust oder aus Sorge über eine zu geringe Rentenzahlung. Dies könnte dazu führen, dass der Darlehensnehmer in Rückzahlungsschwierigkeiten gerät. Ein weitere Ursache für den Darlehensverzicht könnte bereits vorhandenes Wohneigentum sein.

- **Vermögenswirksame Leistungen**
  Der Bezug von vermögenswirksamen Leistungen als Ursache für den Darlehensverzicht deutet darauf hin, dass der Bausparvertrag ausschließlich der Anlage der vermögenswirksamen Leistungen diente. Für vermögenswirksame Leistungen sind nur bestimmte Sparformen zulässig[2]. Neben Fonds und Banksparplänen sind vor allem Bausparverträge durch eine relativ hohe und garantierte Verzinsung mit zusätzlichem Anreiz der Wohnungsbauprämie für Sparer interessant.

- **Geringe Spardauer**
  Bausparkonten mit geringer Spardauer könnten ebenfalls ausschließlich zur Geldanlage dienen. Die häufig sofortige Auffüllung bis zum Mindestanspargrad lässt diesen Schluss zu. Möglicher Grund für dieses Verhalten kann z. B. ein niedriger Außenzins sein.

In der Struktur der Darlehensnehmer sind die Merkmale hohe Wohnungsbauprämie und mittlere, bzw. geringe Bausparsumme auffallend häufig vertreten. Dies lässt folgende Schlussfolgerungen zu:

- **Hohe Wohnungsbauprämie**
  Die Wohnungsbauprämie betrug im zugrunde liegenden Untersuchungsjahr (Jahr 2000) noch 10 % auf maximal 512 Euro für Alleinstehende, bzw. 1.024 Euro Sparzahlungen für Verheiratete. Allerdings darf das zu versteuernde Einkommen bestimmte Grenzwerte nicht überschreiten. Bausparer, die WoP erhalten, liegen also unter besagtem Grenzwert und benötigen daher wahrscheinlich zum Wohnungsbau bzw. Wohnungserwerb das Bauspardarlehen.

- **Mittlere bzw. geringe Bausparsumme**
  Die Bausparsumme ist hier als gering einzustufen (<15.000 Euro). Davon müssen im Tarif 1 nur 40 % eingezahlt werden um die Zuteilung zu erreichen. Die Tilgungsbelastung ist für den Bausparer geringer, da sich der Mindesttilgungssatz prozentual zur Bausparsumme ermittelt. Das Risiko aufgrund der Tilgungsraten in Zahlungsschwierigkeiten zu geraten, ist also geringer. Dies könnte eine Ursache dafür sein, dass das Bauspardarlehen vom Sparer in Anspruch genommen wird.

Die Kombinatorik erlaubt die Bildung von $|\mathcal{P}(4)|=16$ Teilmengen in konjunktiver Normalform, die erfüllt sind, wenn jede Regel der Konjunktion erfüllt ist. Z. B. ist die Regel $(\mathcal{L}_{12}\mathcal{L}_{22})$ genau dann erfüllt, wenn $\mathcal{L}_{12}$ und $\mathcal{L}_{22}$ erfüllt sind. Dabei bezeichnet die leere Menge den Fall, dass das Bausparkonto keine Formel erfüllt. Jedes Bausparkonto wird nun der maximal erfüllten Teilmenge zugeordnet. Die so entstandenen Regeln müssen allerdings noch bewertet werden. Daher muss für jede mögliche Teilmenge,

[2] Nähere Informationen finden sich z. B. im fünften Vermögensbildungsgesetz § 2 unter `http://bundesrecht.juris.de/vermbg_2/index.html`.

die Anzahl der Konten ermittelt werden, die diese erfüllen. Anschließend wird für das Vorhandensein jeder Kombination die Bayessche Wahrscheinlichkeit ermittelt.

Der Ergebnisraum Ω besteht aus allen für die Untersuchung relevanten Konten. Dabei liefern die Mengen D^+ und D^- eine vollständige Zerlegung des Ergebnisraumes. Für D^+ und D^- gilt:

$$\begin{aligned} D^+ \cap D^- &= \emptyset \\ D^+ \cup D^- &= \Omega \end{aligned}$$

Ein Ereignis A bezeichnet in unserem Fall das Vorliegen einer bestimmten Teilmenge aus $\mathcal{P}(\mathcal{L}_{ij})$, $i, j = 1, 2$. Dann gilt für jede Zerlegung und jedes Ereignis A:

$$P(A) = \sum_l P(D^l) \cdot P(A|D^l) \quad l \in \{+, -\}.$$

Dabei bezeichnet $P(D^l)$, $l \in \{+, -\}$ die a priori Verteilung der Darlehensverzichter und Darlehensnehmer im Bestand, $P(A|D^l)$, $l \in \{+, -\}$ entspricht der Wahrscheinlichkeit, dass Ereignis A eintrifft, wenn der Bausparvertrag aus der Menge D^l, $l \in \{+, -\}$ stammt. Daher kann für die Ermittlung der bedingten Wahrscheinlichkeiten die Formel von Bayes verwendet werden. Ist $P(A) > 0$ und gelten die Voraussetzungen der Formel der totalen Wahrscheinlichkeit, so gilt für alle l:

$$P(D^l|A) = \frac{P(D^l) \cdot P(A|D^l)}{\sum_l P(D^l) \cdot P(A|D^l)} \quad l \in \{+, -\}.$$

Für jedes Bausparkonto wird die bedingte Wahrscheinlichkeit ermittelt, dass das Konto aus der Menge D^+ oder D^- stammt. In Tabelle 8.2 sind die ermittelten Wahrscheinlichkeiten für jede Teilmenge aus $\mathcal{P}(\mathcal{L}_{ij})$ dargestellt.

8.1.4 Ergebnisse

Die bedingten Wahrscheinlichkeiten $P(D^+|A)$ liefern uns Schwellwerte, die zur Ermittlung von ROC–Kurven verwendet werden können. Dies ermöglicht die Berechnung verschiedener Werte für Sensitivität und Spezifität. Tabelle 8.3 zeigt Sensitivität und Spezifität der Trainings- und Testmenge für ausgewählte Schwellwerte. Zusätzlich sind die Ergebnisse in Abbildung 8.1 graphisch dargestellt. Die Gesamtquote der Bausparkonten, die von dem Regelwerk überdeckt werden, beträgt in der Trainings- und Testmenge ca. 70 %. Bisherige Analysen gingen davon aus, dass Darlehensverzichte primär vom jeweiligen Außenzins abhängig sind. Dieser Aussage kann sicherlich nicht widersprochen werden, allerdings wurden mit Hilfe des verbandstheoretischen Implikationenmodells zusätzliche Strukturen entdeckt, die auf Darlehensverzichte bzw. Darlehensannahme hindeuten können. Im Hinblick auf die Entwicklung

Vorhandene Regeln des Ereignisses A	$P(D^+\|A)$ (in %)	$P(D^-\|A)$ (in %)
$\mathcal{L}_{11}$	59.41	40.59
$\mathcal{L}_{12}$	19.34	80.66
$\mathcal{L}_{21}$	16.29	83.71
$\mathcal{L}_{22}$	19.79	80.21
$\mathcal{L}_{11}\mathcal{L}_{12}$	55.82	44.18
$\mathcal{L}_{11}\mathcal{L}_{21}$	48.06	51.94
$\mathcal{L}_{11}\mathcal{L}_{22}$	48.77	51.23
$\mathcal{L}_{12}\mathcal{L}_{21}$	24.42	75.58
$\mathcal{L}_{12}\mathcal{L}_{22}$	15.76	84.24
$\mathcal{L}_{21}\mathcal{L}_{22}$	15.47	84.53
$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}$	64.36	35.64
$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{22}$	53.33	46.67
$\mathcal{L}_{11}\mathcal{L}_{21}\mathcal{L}_{22}$	62.20	37.80
$\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22}$	10.59	89.41
$\mathcal{L}_{11}\mathcal{L}_{12}\mathcal{L}_{21}\mathcal{L}_{22}$	57.52	42.48

Tabelle 8.2: Bedingte Wahrscheinlichkeiten für den Darlehensverzicht bzw. die Darlehensnahme im verbandstheoretischen Implikationenmodell

		Trainingsmenge		Testmenge
$P(D^+\|A) >$	Sensitivität	Spezifität	Sensitivität	Spezifität
0.49	0.799	0.579	0.779	0.493
0.50	0.770	0.600	0.740	0.515
0.55	0.682	0.654	0.673	0.572

Tabelle 8.3: Sensitivität und Spezifität der Trainings- und Testmenge für ausgewählte Schwellwerte

des gesamten Darlehenskollektives sind die Ergebnisse von Interesse.
Tabelle 8.3 zeigt, dass ein sensibler Bereich bei $P(D^+|A) > 0.55$ liegt. Hier werden für die Trainings- und Testmenge gleichzeitig die höchsten Werte angenommen. In der Testmenge sinkt die Quote der richtig erkannten Darlehensnehmer allerdings auf 0.57. Trotzdem sind die ermittelten Strukturen aufschlussreich und können unterstützend bei der Klassifikation mitwirken.
Allerdings müsste die Modellierung für verschiedene Zinszeiträume durchgeführt werden, da zur Ermittlung der bedingten Wahrscheinlichkeiten die a priori Wahrscheinlichkeiten über die Verteilung der Darlehensverzichter und Darlehensnehmer im Bestand verwendet wurden.

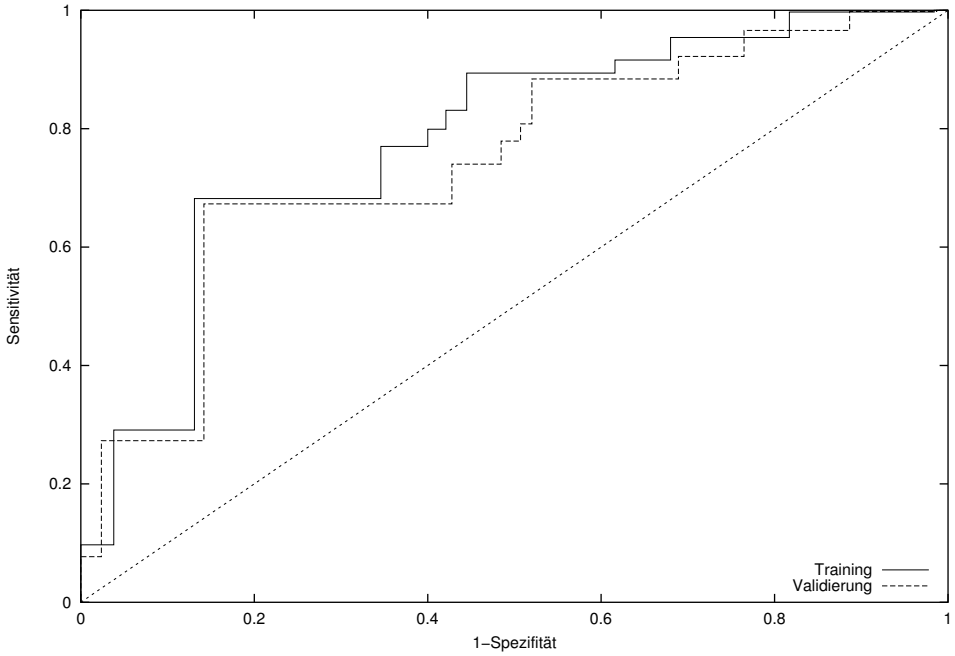


Abbildung 8.1: ROC–Graph der Validierungsbausparkasse für die Klassifikation von Darlehensverzichtern

8.2 Klassifizierung von Kündigern in Sperrfrist

8.2.1 Untersuchungsaufbau

Der Bausparer hat in der Sparphase jederzeit das Recht seinen Bausparvertrag zu kündigen [TCDL02]. Dabei werden zwei Formen der Kündigung unterschieden, zum einen die Kündigung innerhalb der Sperrfrist, zum anderen die Kündigung außerhalb der Sperrfrist, die derzeit sieben Jahre beträgt. Wird ein Bausparvertrag innerhalb von sieben Jahren gekündigt, so muss die erhaltene Wohnungsbauprämie zurückgezahlt werden. Bei Bausparverträgen, die keine Wohnungsbauprämie bezogen haben, ist es unerheblich ob die Kündigung inner- oder außerhalb der Sperrfrist erfolgte.
Es stellt sich die Frage, ob es Merkmale bzw. Merkmalskombinationen gibt, die darauf hindeuten, dass ein Bausparvertrag innerhalb der Sperrfrist gekündigt wird. Mit Sicherheit wird der Bezug von Wohnungsbauprämie eine große Rolle spielen. Eventuell können aber auch zusätzliche Merkmalskombinationen von Bedeutung sein.
Ähnlich wie beim Darlehensverzicht wird davon ausgegangen, dass sich der Bausparer bei einer Kündigung stark am aktuellen Marktzinsniveau orientiert. Liegt das Marktzinsniveau deutlich über dem Sparzins des Bausparvertrages, so ist dies ein Anreiz für den Bausparer seinen Vertrag zu kündigen und das freiwerdende Kapital zu einem höheren Zinssatz anzulegen. Ist der Außenzins hingegen geringer als die Sparzinsen, so entfällt der ökonomische Vorteil für den Bausparer. Um den Zinseinfluss zu verringern

werden, wie bereits bei den Darlehensverzichtern, nur Kündiger in der Sperrfrist des Jahres 2000 betrachtet. Die Menge der nicht gekündigten Bausparverträge besteht dann aus Konten, die sich im Jahr 2000 noch in der Sparphase befinden und älter als sieben Jahre sind (diese können also nicht mehr in der Sperrfrist kündigen) und Konten, die im Jahr 2000 zugeteilt worden sind und daher nicht von ihrem Kündigungsrecht Gebrauch gemacht haben. Zur Datengewinnung wurde erneut das reale Kollektiv der Validierungsbausparkasse (ohne VK/ZK–Verträge) verwendet. Die Auswahlbedingungen lieferten 27.248 Künder in Sperrfrist und 92.361 Nicht–Kündiger in Sperrfrist für das Untersuchungsjahr 2000. Folgende Merkmale wurden zur Erzeugung des Regelwerks verwendet:

- Weiterer Bausparvertrag vorhanden
- Tarif
- Bausparsumme
- Beruf des Bausparers
- WoP–Bezug
- WoP–Höhe
- VL–Bezug
- Lastschrifteinzug vorhanden (LEV)
- Alter des Bausparers

Die Merkmale WoP–Höhe, Alter und Bausparsumme wurden, wie in Unterabschnitt 5.6.2 beschrieben, diskretisiert. Die Merkmale VL–Bezug, WoP–Bezug, Beruf, Tarif, Lastschrifteinzug und weiterer Bausparvertrag vorhanden wurden binär kodiert. Die Daten wurde in eine Trainings- und Testmenge unterteilt. Aus der Trainingsmenge entstanden die Mengen K^+, die das Zielattribut Kündigung in Sperrfrist aufweisen und K^-, die das Zielattribut nicht besitzen, also ihren Bausparvertrag nicht in der Sperrfrist gekündigt haben.
Das verbandstheoretische Implikationenmodell soll nun mit Hilfe eines signifikanten Regelwerks die Kündiger und Nicht–Kündiger klassifizieren. Zudem sollen Kündigerwahrscheinlichkeiten ermittelt werden, die es erlauben, ungesehene Bausparverträge zu kategorisieren.

8.2.2 Ermittlung der Stammbasis und signifikanter Implikationen

Da zwischen dem Erhalt von Wohnungsbauprämie und einer Kündigung in Sperrfrist ein enger Zusammenhang besteht, wurde im Vorfeld eine Analyse bezüglich des Merkmals „Bezug von Wohnungsbauprämie" durchgeführt. Die Analyse der Trainingsdaten

lieferte die Kontingenztafel in Tabelle 8.4. Bei alleiniger Verwendung der Entschei-

	WoP bezogen	keine WoP bezogen	$\sum$
Kündiger	63	1.936	1.999
Nicht–Kündiger	1.555	446	2.001
$\sum$	1.618	2.382	4.000

Tabelle 8.4: Kontingenztabelle zum WoP–Bezug der Kündiger und Nicht–Kündiger in Sperrfrist der Validierungsbausparkasse

dungsregel „WoP–Bezug entspricht der Klasse K^-" und „Kein WoP–Bezug entspricht der Klasse K^+, ergeben sich für Sensitivität und Spezifität die Werte 0.968 und 0.777, d. h. die Verwendung des Merkmals WoP–Bezug liefert bereits ein äußerst hohes Klassifikationsniveau. Es könnte jedoch versucht werden, die Anzahl der False Positives (=Nicht–Kündiger, die keine WoP bezogen haben und fälschlicherweise als Kündiger eingestuft werden) zu verringern. Damit würde sich das Ergebnis der Sensitivität verschlechtern, es würde aber insgesamt ein konstanteres Klassifikationsniveau erreicht. Es wird daher nach Regeln gesucht, die in der Lage sind, die Nicht–Kündiger in Sperrfrist, die keine WoP bezogen haben, zu klassifizieren.

Da die Menge K^+ bereits mit dem Merkmal WoP–Bezug klassifiziert werden kann, wurde nur aus der Menge K^- ein formaler Kontext $\mathbb{K}^-$ erzeugt und die Stammbasis ermittelt. Aus der Stammbasis wurden Implikationen mit erfüllter Prämisse ausgewählt, die den Schwellwert δ=10 % überschreiten. Anschließend wurde ein χ^2–Unabhängigkeitstest für die Implikationen durchgeführt.

8.2.3 Erstellung der Regelsätze

Aufgrund der starken bauspartechnischen Zusammenhänge weicht die Erstellung des Regelsatzes zur Klassifizierung von Kündigern bzw. Nicht–Kündigern vom Standardvorgehen ab. Die erste Analyse zeigte eine enorme Abhängigkeit des Kündigerverhaltens vom WoP–Bezug. Daher werden die Regeln:

- $\mathcal{L}_1$: Kein WoP–Bezug
- $\mathcal{L}_{21}$: WoP–Bezug

direkt zur Klassifikation verwendet. Zur Erlangung einer ausgeglicheren Klassifikationsgüte werden weitere Regeln zur Klassifizierung von Nicht–Kündigern verwendet. Die Signifikanzprüfung lieferte eine Menge $\mathcal{L}_2$ mit signifikanten Implikationen für die Struktur der Menge K^-, die mit Hilfe des Medians in die Mengen $\mathcal{L}_{22}$ und $\mathcal{L}_{23}$ geteilt wurde. Die Teilmengen in disjunktiver Normalform beinhalten folgende Implikationen:

- $\mathcal{L}_{22}$: Stark signifikante Regeln des Kontextes $\mathbb{K}^-$

  VL–Bezug $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden
  BS hoch $\wedge$ VL–Bezug $\wedge$ LEV $\wedge$ Mittleres Alter $\rightarrow$ Tarif 1
  Arbeiter $\wedge$ VL–Bezug $\wedge$ LEV $\rightarrow$ Mittleres Alter
  BS hoch $\wedge$ Angestellter $\wedge$ LEV $\rightarrow$ Mittleres Alter

- $\mathcal{L}_{23}$: Signifikante Regeln des Kontextes $\mathbb{K}^-$

  Tarif 3 $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden
  Tarif 1 $\wedge$ Arbeiter $\wedge$ VL $\rightarrow$ Mittleres Alter
  BS mittel $\wedge$ LEV $\wedge$ alt $\rightarrow$ Weiteres Konto vorhanden
  Rentner $\wedge$ LEV $\rightarrow$ Weiteres Konto vorhanden

Für das Merkmal VL–Bezug gilt ebenfalls eine Sperrfrist. Daher verwundert es nicht, dass das Merkmal signifikant häufiger in der Klasse der Nicht–Kündiger zu finden ist. Daneben sind die Merkmale LEV vorhanden und mittleres Alter in den Implikationen der Nicht–Kündiger anzutreffen. Die Nicht–Kündiger zeichnen sich also durch ein regelmäßiges Sparverhalten mit Lastschrifteinzug aus. Häufig tritt die Kombination VL und LEV in den Implikationen auf.

Die Regeln $\mathcal{L}_1$ und $\mathcal{L}_{21}$ sind disjunkt, d. h. sie können nicht gemeinsam in einem Konto auftreten. Weiterhin können $\mathcal{L}_{22}$ und $\mathcal{L}_{23}$ nicht alleine auftreten, da entweder $\mathcal{L}_1$ oder $\mathcal{L}_{21}$ stets im Konto vorhanden ist. Diese Einschränkungen erlauben letztendlich neun mögliche Kombinationen (einschließlich der $\emptyset$), die in Tabelle 8.5 dargestellt sind. Jedes Bausparkonto wird nun der maximalen erfüllten Teilmenge zugeordnet. Die

$$\begin{array}{cccc} & \emptyset & & \\ & (\mathcal{L}_1) & (\mathcal{L}_{21}) & \\ (\mathcal{L}_1\mathcal{L}_{22}) & (\mathcal{L}_1\mathcal{L}_{23}) & (\mathcal{L}_{21}\mathcal{L}_{22}) & (\mathcal{L}_{21}\mathcal{L}_{23}) \\ & (\mathcal{L}_1\mathcal{L}_{22}\mathcal{L}_{23}) & (\mathcal{L}_{21}\mathcal{L}_{22}\mathcal{L}_{23}) & \end{array}$$

Tabelle 8.5: Mögliche Regelkombinationen zur Klassifizierung von Kündigern und Nicht–Kündigern in Sperrfrist

neun Teilmengen müssen allerdings noch mit Hilfe bedingter Wahrscheinlichkeiten bewertet werden.
Der Ergebnisraum Ω besteht aus allen für die Untersuchung relevanten Konten. Dabei liefern die Mengen K^+ und K^- eine vollständige Zerlegung des Ergebnisraumes. Für K^+ und K^- gilt wiederum $K^+ \cap K^- = \emptyset$ und $K^+ \cup K^- = \Omega$. Ein Ereignis A bezeichnet in unserem Falle dass Vorliegen einer bestimmten Teilmenge aus $\mathcal{P}(\mathcal{L}_{ij})$, $i, j = 1, 2$.

Dann gilt für jede Zerlegung und jedes Ereignis A:

$$P(A) = \sum_l P(K^l) \cdot P(A|K^l) \quad l \in \{+,-\}$$

Dabei bezeichnet $P(K^l)$, $l \in \{+,-\}$ die a priori Verteilung der Kündiger und Nicht–Kündiger im Bestand, $P(A|K^l)$, $l \in \{+,-\}$ entspricht der Wahrscheinlichkeit, dass Ereignis A eintrifft, wenn der Bausparvertrag aus der Menge K^l, $l \in \{+,-\}$ stammt. Daher kann für die Ermittlung bedingter Wahrscheinlichkeiten die Formel von Bayes verwendet werden. Ist $P(A) > 0$ und gelten die Voraussetzungen der Formel der totalen Wahrscheinlichkeit, so gilt für alle l:

$$P(K^l|A) = \frac{P(K^l) \cdot P(A|K^l)}{\sum_l P(K^l) \cdot P(A|K^l)} \quad l \in \{+,-\}$$

Für jedes Bausparkonto kann die bedingte Wahrscheinlichkeit ermittelt werden, dass das Konto aus der Menge K^+ oder K^- stammt. Mit Hilfe der Bayesschen Formel können bedingte Wahrscheinlichkeiten für alle Teilmengen ermittelt werden, die in Tabelle 8.6 dargestellt sind.

Vorhandene Regeln des Ereignisses A	$P(K^+\|A)$ (in %)	$P(K^-\|A)$ (in %)
$\mathcal{L}_1$	74.12	25.88
$\mathcal{L}_{21}$	2.77	97.23
$\mathcal{L}_1\mathcal{L}_{22}$	49.45	50.55
$\mathcal{L}_1\mathcal{L}_{23}$	60.49	39.51
$\mathcal{L}_{21}\mathcal{L}_{22}$	0.20	99.80
$\mathcal{L}_{21}\mathcal{L}_{23}$	1.40	98.60
$\mathcal{L}_1\mathcal{L}_{22}\mathcal{L}_{23}$	36.64	63.36
$\mathcal{L}_{21}\mathcal{L}_{22}\mathcal{L}_{23}$	0.35	99.65

Tabelle 8.6: Bedingte Wahrscheinlichkeiten für die Kündigung und Nicht–Kündigung in Sperrfrist im verbandstheoretischen Implikationenmodell

8.2.4 Ergebnisse

Die bedingten Wahrscheinlichkeiten ermöglichen eine Analyse der Klassifikationsergebnisse mit Hilfe von ROC–Graphen. Dazu werden Sensitivität und Spezifität an den jeweiligen Schwellwerten ermittelt. Ein Vorteil der Verwendung des disjunkten Merkmals WoP–Bezug bzw. kein WoP–Bezug liegt darin, dass alle Konten bewertbar sind, da jeder Bausparvertrag entweder das eine oder das andere Merkmal erfüllen muss. Die Gesamtquote der klassifizierten Konten liegt daher bei 100 %.

		Trainingsmenge		Testmenge
$P(K^+\|A) >$	Sensitivität	Spezifität	Sensitivität	Spezifität
0.45	0.942	0.800	0.942	0.788
0.55	0.866	0.839	0.866	0.827
0.60	0.866	0.859	0.866	0.849
0.70	0.805	0.859	0.805	0.849

Tabelle 8.7: Sensitivität und Spezifität der Trainings- und Testmenge für ausgewählte Schwellwerte

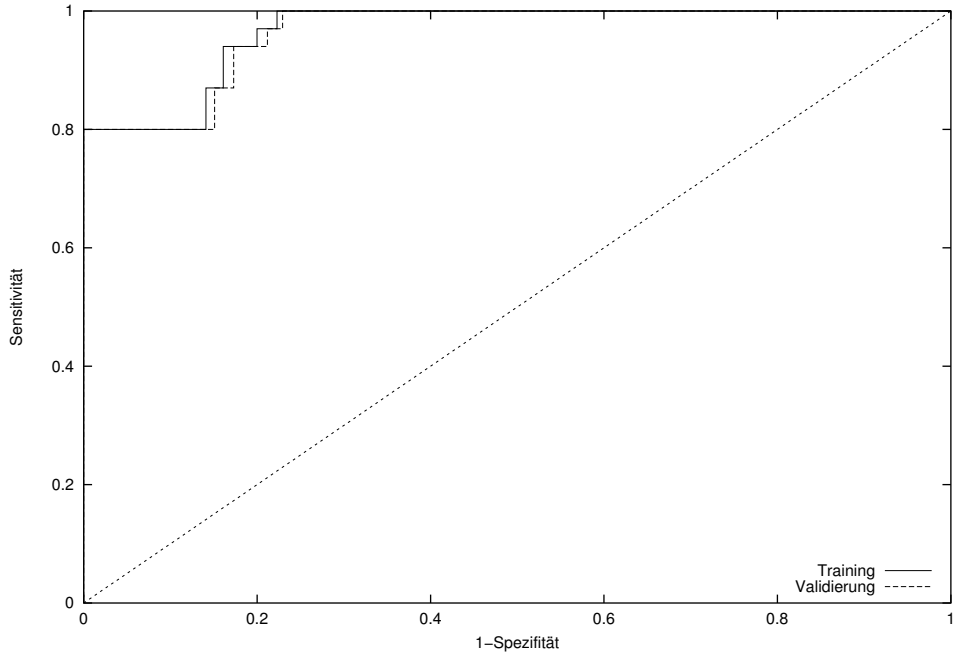


Abbildung 8.2: ROC–Graph der Validierungsbausparkasse für die Klassifikation von Kündigern in Sperrfrist

Kapitel 9

Zusammenfassung und Ausblick

Im Hinblick auf die Umsetzung von Basel II sind interne Bonitätsbeurteilungen für Banken und Bausparkassen von enormer Bedeutung, da häufig nicht auf externe Ratings zurückgegriffen werden kann. Daher wurde in dieser Arbeit ein verbandstheoretisches Klassifikationsmodell entwickelt, welches eine natürliche Einteilung der Kreditnehmer in Bonitätsklassen mit zugehöriger Ausfallwahrscheinlichkeit liefert. Da bei Privatkrediten keine Bilanz- oder Unternehmenskennzahlen vorliegen, stützt sich das Modell bei der Erstellung von logischen Regeln auf vertragsimmanente sowie persönliche Merkmale. Mit Hilfe der logischen Regeln werden die Bauspardarlehen bewertet und klassifiziert. Die Verwendung von logischen Regeln erleichtert zudem die Interpretation der Ergebnisse.
Logische Ansätze verlangen stets eine Diskretisierung der Eingangsdaten, die im verbandstheoretischen Implikationenmodell mit Hilfe von bauspartechnischem Expertenwissen durchgeführt wird. Zudem wird auf eine mögliche Verwendung der ermittelten Ratingklassen im Rahmen von IRB–Ansätzen eingegangen. Zur Bewertung der Klassifikationsgüte wird die Problemstellung mit Hilfe weiterer Klassifikationsverfahren bearbeitet und die Ergebnisse mit denen des verbandstheoretischen Implikationenmodells verglichen. Dies sind im Einzelnen folgende Modelle: Neuronale Netze, Entscheidungsbäume und das logische Modell von Truemper. Zum Vergleich der logischen Modelle wurde mit einer einheitlichen Diskretisierung der Eingangsdaten gearbeitet. Zur Prüfung der Generalisierungsfähigkeit aller Modelle wurden die Ergebnisse ohne weitere Anpassung auf die Validierungsbausparkasse angewendet. Dies ist für die Bausparkassen von hohem Interesse, da durch das bausparspezifische Spezialgeschäft große Ähnlichkeiten in den Portfolios existieren. Damit könnte eine bausparkassenspezifische Modellierung in den einzelnen Häusern entfallen. Zusätzlich werden die Modelle quantitativ und qualitativ verglichen, um Gemeinsamkeiten und Unterschiede in den Modellen herauszuarbeiten. In einem letzten Schritt wurde die Anwendung des verbandstheoretischen Implikationenmodells auf weitere bauspartechnische Fragestellungen untersucht. Dabei handelte es sich um die Klassifikation von Darlehensverzichtern und Darlehensnehmern, sowie die Bewertung von Kündigern und Nicht–Kündigern in Sperrfrist. Diese Fragestellungen sind unter anderem für die Kollektiv-

entwicklung der Bausparkassen von hohem Interesse.

Die Analyse der Klassifikationsergebnisse zeigte enorme Unterschiede bei den jeweiligen Modellen. Während die neuronalen Netze für die Ausgangsbausparkasse die besten Ergebnisse lieferten, verschlechterten sie sich bei der Generalisierung drastisch. Allerdings sind neuronale Netze stets in der Lage alle Darlehenskonten zu bewerten. Das verbandstheoretische Implikationenmodell und die anderen logischen Modelle lieferten teilweise nur geringfügig niedrigere Ergebnisse als die neuronalen Netzen, bieten aber durch ihre bessere Generalisierungsfähigkeit ein breiteres Anwendungsspektrum.
Ein Vorteil des verbandstheoretischen Implikationenmodells liegt in der Erzeugung eines für den Anwender überschaubaren und nachvollziehbaren Regelwerks. Zudem liefert das Modell aufgrund der Zusammenfassung der Regeln auf natürliche Weise eine Zerlegung des Darlehenbestandes in unterschiedliche Bonitätsklassen. Die Anzahl der Bonitätsklassen kann dabei über den Grad der Zusammenfassung der Regeln gesteuert werden. Eine empirische Bewertung erlaubt die Ermittlung von Kreditausfallwahrscheinlichkeiten für die Ratingklassen. Durch Angabe der zugrunde liegenden Regeln sind die Ratingklassen klar definiert und ordnen Kreditnehmer mit gleichen Merkmalen derselben Bonitätsklasse zu. Auch die hohe Generalisierungsfähigkeit und die erfolgreiche Anwendung auf weitere bauspartechnische Fragestellungen sind von Vorteil.

Das verbandstheoretische Implikationenmodell besitzt noch einige Schwächen, die einen Ausgangspunkt für Weiterentwicklungen bieten. Schwächen liegen vor allem in der variablen Anzahl und einer geeigneten Zusammenfassung der Regeln. Hier wurden vereinfachte Annahmen getroffen, die unter anderem mit der Anzahl der vorhandenen Bauspardarlehen in den Stichproben begründet wurden. Daneben müssen die gewonnenen Bonitätsklassen noch um eine Klasse für bereits ausgefallene Darlehen sowie eine Klasse noch nicht bewerteter Darlehen erweitert werden. Allerdings fehlt im zweiten Fall eine charakteristische Beschreibung dieser Klasse und somit auch die zugehörige Kreditausfallwahrscheinlichkeit. Bei Stichproben mit geringem Umfang könnte die empirische Ermittlung der Ausfallwahrscheinlichkeiten zudem Probleme bereiten, da nicht notwendigerweise alle Teilmengen der Potenzmenge belegt sein müssen.

Trotz den erwähnten Nachteilen und Schwächen lieferte das verbandstheoretische Implikationenmodell für alle untersuchten Problemstellungen relativ robuste Ergebnisse. Seine natürliche Klassenzerlegung mit empirisch ermittelten Ausfallwahrscheinlichkeiten legt zudem die Verwendung des Modells im Rahmen von IRB–Ansätzen nahe. Zur Klassenerweiterung könnte das verbandstheoretische Implikationenmodell in Kombination mit anderen Modellen eingesetzt werden. Aufgrund der hohen Generalisierungsfähigkeit bietet sich eine Anwendung des Modells zum Beispiel im Bereich der Privatkundenkredite an.

Literaturverzeichnis

[AHK97] ANDRESS, H.-J., J.A. HAGENAARS und S. KÜHNEL: *Analyse von Tabellen und kategorialen Daten*. Springer, Berlin, 1997.

[And97] ANDERS, U.: *Statistische neuronale Netze*. Dissertation, Universität Fridericiana zu Karlsruhe, Karlsruhe, 1997.

[Bas04] BASELER AUSSCHUSS FÜR BANKENAUFSICHT: *Internationale Konvergenz der Kapitalmessung und Eigenkapitalanforderung*. Basel, 2004.

[Bau91] BAUER, H.: *Wahrscheinlichkeitstheorie*. Walter de Gruyter, Berlin, New York, 4. Auflage, 1991.

[BB96] BAMBERG, G. und F. BAUR: *Statistik*. Oldenbourg, München, Wien, 9. Auflage, 1996.

[BCK04] BREZSKI, E., C. CLAUSSEN und H.-M. KORTH: *Rating – Basel II und die Folgen*. Richard Boorberg Verlag, Stuttgart, 2004.

[BEP+86] BACKHAUS, K., B. ERICHSON, W. PLINKE, CHR. SCHUCHARD-FICHER und R. WEIBER: *Multivariate Analysemethoden*. Springer, Berlin, Heidelberg, 4. Auflage, 1986.

[BGMT04] BARTNIKOWSKI, S., M. GRANBERRY, J. MUGAN und K. TRUEMPER: *Transformation of Rational and Set Data to Logic Data*. In: TRIANTAPHYLLOU, E. und G. FELICI (Herausgeber): *Data Mining and Knowledge Discovery Approaches Based on Rule Induction Techniques*, Kapitel 7. Kluwer Academic Publisher, New York, 2004.

[BLK06] BERENSON, M., D. LEVINE und T. KREHBIEL: *Basic Business Statistics*. Pearson Education, New Jersey, 10. Auflage, 2006.

[Boe03] BOECKH, F.: *Fit für Basel II: Bausparkassen mit integrierten Score–Karten*. Immobilien & Finanzierung, (12):412–414, 2003.

[CAN98] CAOUETTE, J. B., E. I. ALTMAN und P. NARAYANAN: *Managing Credit Risk – The Next Great Financial Challenge*. John Wiley & Sons, Inc., Canada, 1998.

[CDEL05] CLUSE, M., A. DERNBACH, J. ENGELS und P. LELLMANN: *Einführung in Basel II*. In: DELOITTE (Herausgeber): *Basel II – Handbuch zur praktischen Umsetzung des neuen Bankenaufsichtsrechts*, Seiten 19–44. Erich Schmidt Verlag, Berlin, 2005.

[Che05] CHEVALIER, T.: *Ein Risikomodell für Bausparkollektive*. Dissertation, Universität zu Köln, Köln, 2005.

[CL03] CONNOR-LINTON, J.: *Chi Square Tutorial*. Univercity of Georgetown, 2003. `http://www.georgetown.edu/faculty/ballc/webtools/web_chi_tut.html`.

[CM03] CASPARD, N. und B. MONJARDET: *The lattices of closure systems, closure operators, and implicational systems on a finite set: a survey*. Discrete Applied Mathematics, 127(2):241–269, 2003.

[Cre04] CREDITSUISSE: *Basel II – Meilenstein der Bankenregulierung*. Credit Suisse Economic Research, Zürich, 2004.

[Cre06] CREDITREFORM: *Insolvenzen Neugründungen Löschungen 1. Halbjahr 2006*. Creditreform Wirtschaftsforschung, Neuss, 2006.

[Deu04] DEUTSCHE BUNDESBANK: *Neue Eigenkapitalanforderungen für Kreditinstitute (Basel II). Monatsbericht September 2004*. Frankfurt, 2004.

[DG86] DUQUENNE, V. und J.-L. GUIGUES: *Famille minimale d'implications informatives resultant d'un tableau de donnees binaires*. Mathématiques et Sciences Humaines, 24(95):5–18, 1986.

[DHS01] DUDA, R., P. HART und D. STORK: *Pattern Classification*. John Wiley & Sons, Inc., Canada, 2. Auflage, 2001.

[EFM04] EVERDING, D., P. FAKLER und A. MUEMKEN: *Persönliche Kommunikation*, 2004.

[EFT92] EBBINGHAUS, H.-D., J. FLUM und W. THOMAS: *Einführung in die mathematische Logik*. BI Wissenschaftsverlag, Mannheim/Leipzig/Wien/Zürich, 3. Auflage, 1992.

[Faw03] FAWCETT, T.: *ROC Graphs: Notes and Practical Considerations for Researchers*. Technical report, HP Laboratories, 2003.

[FF92] FAHRMEIR, L. und H. FROST: *On Stepwise Variable Selection in Generalized Linear Regression and Time Series Models*. Computational Statistics, 92(2):137–154, 1992.

[Füs95] FÜSER, K.: *Neuronale Netze in der Finanzwirtschaft: innovative Konzepte und Einsatzmöglichkeiten*. Gabler Verlag, Wiesbaden, 1995.

[FT97] FAHRMEIR, L. und G. TUTZ: *Multivariate Statistical Modelling based on generalized linear Models*. Springer, München, 2. Auflage, 1997.

[FT02] FELICI, G. und K. TRUEMPER: *A MINSAT Approach for Learning in Logic Domains.* INFORMS Journal on Computing, 14(1):20–36, 2002.

[Gan99] GANTER, B.: *Attribute exploration with background knowledge*. Theoretical Computer Science, 217(2):215–233, 1999.

[Gan00] GANTER, B.: *Begriffe und Implikationen*. In: STUMME, G. und R. WILLE (Herausgeber): *Begriffliche Wissensverarbeitung: Methoden und Anwendungen*, Seiten 1–24. Springer, 2000.

[GK00] GANTER, B. und S.O. KUZNETSOV: *Formalizing Hypotheses with Concepts*. In: GANTER, B. und G.W. MINEAU (Herausgeber): *Conceptual Structures: Logical, Linguistic, and Computational Issues*, Band 1867 der Reihe *Lecture Notes in Computer Science*, Seiten 342–356. Springer, 2000.

[GK01] GANTER, B. und S.O. KUZNETSOV: *Pattern Structures and Their Projections*. In: DELUGACH, H.S. und B. STUMME (Herausgeber): *Conceptual Structures: Broadening the Base*, Band 2120 der Reihe *Lecture Notes in Computer Science*, Seiten 129–142. Springer, 2001.

[GW96] GANTER, B. und R. WILLE: *Formale Begriffsanalyse: Mathematische Grundlagen*. Springer, Berlin, Heidelberg, 1996.

[HHK00] HÄRDLE, W., Z. HLAVKA und S. KLINKE: *XploRe Application Guide*. Springer, Berlin, 2000.

[HM82] HANLEY, J.A. und B.J. MCNEIL: *The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve*. Radiology, 143(1):29–36, 1982.

[HS93] HASSIBI, B. und D. G. STORK: *Second Order Derivatives for Network Pruning: Optimal Brain Surgeon*. In: HANSON, S.J., J.D. COWAN und C. LEE GILES (Herausgeber): *Advances in Neural Information Processing Systems 5*, Seiten 164–171. Morgan Kaufmann, San Mateo, CA, 1993.

[HW91] HEINEMANN, B. und K. WEIHRAUCH: *Logik für Informatiker*. B.G. Teubner, Stuttgart, 1991.

[JN06] JANSSEN, P. und L. NOURINE: *Minimum implicational basis for ∧–semidistributive lattices*. Information Processing Letters, 99(5):199–202, 2006.

[Kad06] KADERALI, L.: *A Hierarchical Bayesian Approach to Regression and its Application to Predicting Survival Times in Cancer*. Dissertation, Universität zu Köln, Köln, 2006.

[Kön03] KÖNIG, R.: *Endliche Hüllensysteme und ihre Implikationenbasen*. Séminaire Lotharingien de Combinatoire, 49(B49g):46 pp, 2003.

[KO02] KUZNETSOV, S.O. und S.A. OBIEDKOV: *Comparing Performance of Algorithms for Generating Concept Lattices*. Journal of Experimental and Theoretical Artificial Intelligence, 14(2–3):189–216, 2002.

[Kre03] KRENGEL, U.: *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Vieweg, Wiesbaden, 2003.

[KS00] KAISER, U. und A. SZCZESNY: *Einfache ökonometrische Verfahren für die Kreditrisikomessung: Logit– und Probit–Modelle*, Band 61 der Reihe *Working Paper Series: Finance & Accounting*. Johann Wolfgang Goethe Universität Frankfurt am Main Fachbereich Wirtschaftswissenschaften, Frankfurt/Main, 2000.

[Lau05] LAUX, H.: *Die Bausparfinanzierung*. Verlag Recht und Wirtschaft, Frankfurt am Main, 2005.

[LDS90] LECUN, Y., J. DENKER und S. SOLLA: *Optimal Brain Damage*. In: TOURETZKY, D.S. (Herausgeber): *Advances in Neural Information Processing Systems 2*, Seiten 598–605. Morgan Kaufmann, San Mateo, CA, 1990.

[Lin99] LINDIG, C.: *Algorithmen zur Begriffsanalyse und ihre Anwendung bei Softwarebibliotheken*. Dissertation, Technische Universität Braunschweig, Braunschweig, 1999.

[LMGR00] LEHN, J., T. MÜLLER-GRONBACH und S. RETTIG: *Einführung in die deskreptive Statistik*. B.G. Teubner, Stuttgart, Leipzig, 2000.

[Lug01] LUGER, G. F.: *Künstliche Intelligenz – Strategien zur Lösung komplexer Probleme*. Pearson Studium, München, 4. Auflage, 2001.

[LW00] LEHN, J. und H. WEGMANN: *Einführung in die Statistik*. B.G. Teubner, Stuttgart, Leipzig, 2000.

[MA04] MEUSEL, S.G. und S. ASCHENBRENNER–VON DAHLEN: *Entwicklungen und Parallelen von Basel II & Solvency II*. OR News, (22):5–11, 2004.

[MD89] MONTANA, D. J. und L. DAVIS: *Training Feedforward Neural Networks Using Genetic Algorithms.* In: *Proceedings of the International Joint Conference on Artificial Intelligence*, Seiten 762–767, 1989.

[Meu03] MEUSEL, S.G.: *Bepreisung von Kreditrisiken nach Basel II*. OR News, (18):5–9, 2003.

[Mit97] MITCHELL, T.: *Machine Learning*. WCB McGraw–Hill, Boston, Massachusetts, 1997.

[MP04] MACSKASSY, S.A. und F.J. PROVOST: *Confidence Bands for ROC Curves: Methods and an Empirical Study.* In: *Proceedings of the First Workshop on ROC Analysis in Artificial Intelligence*, Seiten 61–70, 2004.

[MR00] MÜLLER, M. und B. RÖNZ: *Credit Scoring using Semiparametric Methods*. In: FRANKE, J., W. HÄRDLE und W. STAHL (Herausgeber): *Measuring Risk in Complex Stochastic Systems*, Band 147 der Reihe *Lecture Notes in Statistics*, Kapitel 5, Seiten 83–96. Springer, Berlin, 2000.

[MT04] MUGAN, J. und K. TRUEMPER: *Discretization of Rational Data*. In: FELICI, G. und C. VERCELLIS (Herausgeber): *Proceedings of Mathematical Methods for Learning 2004*, Berlin, 2004. IGI Publisher Group.

[Neu98] NEUHAUS, M.: *Modernes Kreditportfolio–Management – Neue Herausforderung durch Anwendung quantitativer Methoden*. Swiss Banking School, Bern, Stuttgart, Wien, 1998.

[OU02] OEHLER, A. und M. UNSER: *Finanzwirtschaftliches Risikomanagement*. Springer, Berlin, 2. Auflage, 2002.

[PD03] PROVOST, F.J. und P. DOMINGOS: *Tree Induction for Probability–Based Ranking.* Machine Learning, 52(3):199–215, 2003.

[Qui93] QUINLAN, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Rem01] REMSHAGEN, A.: *Learning for SAT and MINSAT, and algorithms for quantified SAT and MINSAT*. Dissertation, Univercity of Texas at Dallas, Dallas, 2001.

[RMS90] RITTER, H., T. MARTINETZ und K. SCHULTEN: *Neuronale Netze – Eine Einführung in die Neuroinformatik selbstorganisierender Netzwerke*. Addison–Wesley Publishing Company, Bonn, München, 2. Auflage, 1990.

[RN04] RUSSELL, S. und P. NORVIG: *Künstliche Intelligenz – Ein moderner Ansatz*. Pearson Studium, München, 2. Auflage, 2004.

[Sac02] SACHS, L.: *Angewandte Statistik*. Springer, Berlin, Heidelberg, 10. Auflage, 2002.

[Sch02] SCHRÖDER, M.: *Finanzmarkt–Ökonometrie: Basistechniken, Fortgeschrittene Verfahren, Prognosemodelle*. Schäffer–Poeschel Verlag, Stuttgart, 2002.

[Sch03] SCHEULE, H.: *Prognose von Kreditausfallwahrscheinlichkeiten*. Dissertation, Universität Regensburg, Regensburg, 2003.

[SR94] SCHMIDT VON RHEIN, A. und H. REHKUGLER: *KNN zur Kreditwürdigkeitsprüfung bei Privatkundenkrediten*. In: REHKUGLER, H. und H.G. ZIMMERMANN (Herausgeber): *Neuronale Netze in der Ökonomie*, Seiten 491–545. Vahlen, München, 1994.

[SS94] SEVERINI, T.A und J.G STANISWALIS: *Quasi–likelihood Estimation in Semiparametric Models*. Journal of the American Statistical Association, 89(426):501–511, 1994.

[Stu02] STUMME, G.: *Efficient Data Mining Based on Formal Concept Analysis.* In: *Database and Expert Systems Applications : 13th International Conference, DEXA 2002*, Seiten 534–546, 2002.

[SW49] SHANNON, C.E. und W. WEAVER: *The Mathematical Theory of Communication*. Univercity of Illinois Press, Urbana, Illinois, 1949.

[TCDL02] THOMAS, W., R. CONRADI, B. DIETRICH und H. J. LINDNER: *Bausparkassen–Fachbuch 2002/2003*. Bundesgeschäftsstelle Landesbausparkassen, Berlin, 2002.

[Tru04] TRUEMPER, K.: *Design of Logic–based Intelligent Systems*. John Wiley & Sons, Inc., New Jersey, 2004.

[Van96] VANNAHME, I.: *Clusteralgorithmen zur mathematischen Simulation von Bausparkollektiven*. Dissertation, Universität zu Köln, Köln, 1996.

[Var97] VARNHOLT, B.: *Modernes Kreditrisikomanagement*. Verlag Neue Zürcher Zeitung, Zürich, 1997.

[WBE04] WILKENS, M., R. BAULE und O. ENTROP: *IRB–Ansatz in Basel II – die Behandlung erwarteter Verluste*. Zeitschrift für das gesamte Kreditwesen, 57(14):734–737, 2004.

[WEV01] WILKENS, M., O. ENTROP und J. VÖLKER: *Strukturen und Methoden von Basel II – Grundlegende Veränderungen der Bankenaufsicht*. Zeitschrift für das gesamte Kreditwesen, 54(4):187–193, 2001.

[Wil94] WILD, M.: *A Theory of Finite Closure Spaces Based on Implications*. Advances in Mathematics, 108(1):118–139, 1994.

[WN70] WITTING, H. und G. NÖLLE: *Angewandte Mathematische Statistik*. B.G. Teubner, Stuttgart, 1970.

[Zeh01] ZEHNDER, A. J.: *Basel II und die Bausparkassen*. Immobilien & Finanzierung, (7/8):253–255, 2001.

[Zic91] ZICKWOLFF, M.: *Rule Exploration: First Order Logic in Formal Concept Analysis*. Dissertation, Technische Universität Darmstadt, Darmstadt, 1991.

[Zim94] ZIMMERMANN, H.G.: *Neuronale Netze als Entscheidungskalkül*. In: REHKUGLER, H. und H.G. ZIMMERMANN (Herausgeber): *Neuronale Netze in der Ökonomie*, Seiten 1–87. Vahlen, München, 1994.

Index

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie – abgesehen von unten angegebenen Teilpublikationen – noch nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. R. Schrader betreut worden.

Teilpublikationen:
keine

Lebenslauf

Persönliche Daten

Name	Petra Fakler
Adresse	Landmannstr. 34, 50825 Köln
geboren am	08. Juni1973 in Memmingen
Staatsangehörigkeit	deutsch
Familienstand	unverheiratet

Schulbildung

1980–1984	Grundschule Berkheim
1984–1990	Realschule Erolzheim
1993–1995	Berufsoberschule Memmingen
1995	Allgemeine Hochschulreife

Studium

09/1995–05/2001	Studium der Mathematik und Sport für das gymnasiale Lehramt an der Technischen Universität Darmstadt
05/2001	Erstes Staatsexamen für das gymnasiale Lehramt

Berufstätigkeit

09/1990–01/1993	Ausbildung zur Bankkauffrau bei der Raiffeisenbank Illertal eG
02/1993–08/1993	Tätigkeit als Bankkauffrau bei der Raiffeisenbank Illertal eG
Seit 06/2001	Wissenschaftliche Mitarbeiterin am Mathematischen Institut/Zentrum für angewandte Informatik Köln (ZAIK), Universität zu Köln in einem Drittmittelprojekt mit den Landesbausparkassen

Köln, im Februar 2007